Otmar Scherzer
Editor

Handbook of Mathematical Methods in Imaging

Second Edition

Springer Reference

Handbook of Mathematical Methods in Imaging

Otmar Scherzer
Editor

Handbook of Mathematical Methods in Imaging

Second Edition

With 472 Figures and 23 Tables

Springer Reference

Editor
Otmar Scherzer
Computational Science Center
University of Vienna
Vienna, Austria

RICAM
Austrian Academy of Sciences
Linz, Austria

ISBN 978-1-4939-0789-2 ISBN 978-1-4939-0790-8 (eBook)
ISBN 978-1-4939-0791-5 (print and electronic bundle)
DOI 10.1007/978-1-4939-0790-8

Library of Congress Control Number: 2015936285

Mathematics Subject Classification: 65N22, 65R32, 62H35, 68U10

Springer New York Heidelberg Dordrecht London
© Springer Science+Business Media New York 2011, 2015
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Printed on acid-free paper

Springer Science+Business Media LLC New York is part of Springer Science+Business Media (www.springer.com)

Preface

Today, *computer imaging* covers various aspects of *data filtering*, *pattern recognition*, *feature extraction*, *computer aided design*, and *computer aided inspection* and *diagnosis*.

Pillars of the field of computer imaging are advanced, stable, and reliable algorithms. In addition, feasibility analysis is required to evaluate practical relevance of the methods. To put these pillars on solid grounds, a significant amount of mathematical tools are required. This handbook makes a humble attempt to provide a survey of such mathematical tools.

We had the vision that this imaging handbook should contain individual chapters which can serve as toolboxes, which, when aligned, form background material for complete applied imaging problems. Therefore it should also give an impression on the broad mathematical knowledge required to solve industrial and applied research applications: The image formation process, very frequently, is assigned to the inverse problems community, which is prominently represented in this handbook. The subsequent step is image analysis. Nowadays, advanced Image Analysis, and Image Processing in general, uses sophisticated methods from Geometry, Differential Geometry, Convex Analysis, Numerical Analysis, to mention just a few. In fact, by the rapid advance of Imaging, the mathematical areas have been pushed forward heavily, and raised their impact in application sciences.

Second Edition

My special thanks go to all individual authors for their valuable contributions and the referees for their help in improving the contributions and making detailed comments. My sincere thanks go to the Springer's editors and staff, Marc Strauss, Annalea Manalili, Michael Hermann, and Saskia Ellis for their patience and their constant support and encouragement over the last two years. My thanks also goes to Vaishali Damle, who initialized the second edition of this handbook.

Finally, I would like to encourage the readers to submit suggestions regarding this handbook's content. For a project of this size, it is likely that essential topics are missed. In a rapidly evolving area like Imaging it is likely that new areas will

appear in a very short time and should be added to this handbook, as well as recent development enforce modifications of existing contributions. We are committed to issuing periodic updates and we look forward to the feedback from the community.

Otmar Scherzer
Computational Science Center
University of Vienna, Austria
and
RICAM
Austrian Academy of Sciences
Linz, Austria

About the Editor

Otmar Scherzer was born on June 10, 1964, in Vöcklabruck, Austria. He studied technical mathematics at the University of Linz, Austria, and received his Diploma in Mathematics in 1987. He received his Ph.D. in 1990 and his habilitation in 1995 from the same university. During 1995 and 1996, he visited Texas A&M University and the University of Delaware in USA. From 1999 to 2001, he held professorships at the Ludwig Maximilian University, Munich, and the University of Bayreuth, Germany. Otmar then joined the University of Innsbruck, Austria, where he served as full professor of Applied and Computer Oriented Mathematics, from 2001 to 2009. In 2009, he accepted a position at the University of Vienna, where he currently heads the Computational Science Center which was created upon his appointment. More information about the center can be found at: http://www.csc.univie.ac.at/. He is also group leader of the Inverse Problems and Imaging groups at the Radon Institute of Computational and Applied Mathematics in Linz, Austria. More information on the institute can be found at: http://www.ricam.oeaw.ac.at/.

Otmar's research interests include inverse problems, in particular photoacoustic imaging, regularization, image processing, and PDEs. He is a prolific researcher, with more than a 170 research articles published in several well-respected journals. He is the coauthor of two monographs and had coedited 7 volumes, including this handbook, and has served on the editorial board of many prominent journals.

Contents

Volume 1

Part I Inverse Problems – Methods **1**

Linear Inverse Problems 3
Charles Groetsch

Large-Scale Inverse Problems in Imaging 47
Julianne Chung, Sarah Knepper, and James G. Nagy

Regularization Methods for Ill-Posed Problems 91
Jin Cheng and Bernd Hofmann

Distance Measures and Applications to Multimodal Variational Imaging 125
Christiane Pöschl and Otmar Scherzer

Energy Minimization Methods 157
Mila Nikolova

Compressive Sensing 205
Massimo Fornasier and Holger Rauhut

Duality and Convex Programming 257
Jonathan M. Borwein and D. Russell Luke

EM Algorithms 305
Charles Byrne and Paul P.B. Eggermont

EM Algorithms from a Non-stochastic Perspective 389
Charles Byrne

Iterative Solution Methods 431
Martin Burger, Barbara Kaltenbacher, and Andreas Neubauer

Level Set Methods for Structural Inversion and Image Reconstruction .. 471
Oliver Dorn and Dominique Lesselier

Volume 2

Part II Inverse Problems – Case Examples **533**

Expansion Methods 535
Habib Ammari and Hyeonbae Kang

Sampling Methods 591
Martin Hanke-Bourgeois and Andreas Kirsch

Inverse Scattering 649
David Colton and Rainer Kress

Electrical Impedance Tomography 701
Andy Adler, Romina Gaburro, and William Lionheart

Synthetic Aperture Radar Imaging 763
Margaret Cheney and Brett Borden

Tomography 801
Gabor T. Herman

Microlocal Analysis in Tomography 847
Venkateswaran P. Krishnan and Eric Todd Quinto

Mathematical Methods in PET and SPECT Imaging 903
Athanasios S. Fokas and George A. Kastis

Mathematics of Electron Tomography 937
Ozan Öktem

Optical Imaging 1033
Simon R. Arridge, Jari P. Kaipio, Ville Kolehmainen, and Tanja Tarvainen

Photoacoustic and Thermoacoustic Tomography: Image Formation Principles 1081
Kun Wang and Mark A. Anastasio

Mathematics of Photoacoustic and Thermoacoustic Tomography 1117
Peter Kuchment and Leonid Kunyansky

Mathematical Methods of Optical Coherence Tomography 1169
Peter Elbau, Leonidas Mindrinos, and Otmar Scherzer

Wave Phenomena 1205
Matti Lassas, Mikko Salo, and Gunther Uhlmann

Sonic Imaging 1253
Frank Natterer

Imaging in Random Media 1279
Liliana Borcea

Volume 3

Part III Image Restoration and Analysis 1341

Statistical Methods in Imaging 1343
Daniela Calvetti and Erkki Somersalo

Supervised Learning by Support Vector Machines 1393
Gabriele Steidl

Total Variation in Imaging 1455
V. Caselles, A. Chambolle, and M. Novaga

Numerical Methods and Applications in Total Variation Image Restoration 1501
Raymond Chan, Tony F. Chan, and Andy Yip

Mumford and Shah Model and Its Applications to Image Segmentation and Image Restoration 1539
Leah Bar, Tony F. Chan, Ginmo Chung, Miyoun Jung, Nahum Kiryati, Nir Sochen, and Luminita A. Vese

Local Smoothing Neighborhood Filters 1599
Jean-Michel Morel, Antoni Buades, and Tomeu Coll

Neighborhood Filters and the Recovery of 3D Information 1645
Julie Digne, Mariella Dimiccoli, Neus Sabater, and Philippe Salembier

Splines and Multiresolution Analysis 1675
Brigitte Forster

Gabor Analysis for Imaging 1717
Ole Christensen, Hans G. Feichtinger, and Stephan Paukner

Shape Spaces 1759
Alain Trouvé and Laurent Younes

Variational Methods in Shape Analysis 1819
Martin Rumpf and Benedikt Wirth

Manifold Intrinsic Similarity 1859
Alexander M. Bronstein and Michael M. Bronstein

Image Segmentation with Shape Priors: Explicit Versus Implicit Representations 1909
Daniel Cremers

Optical Flow 1945
Florian Becker, Stefania Petra, and Christoph Schnörr

Non-linear Image Registration 2005
Lars Ruthotto and Jan Modersitzki

Starlet Transform in Astronomical Data Processing 2053
Jean-Luc Starck, Fionn Murtagh, and Mario Bertero

Differential Methods for Multi-dimensional Visual Data Analysis 2099
Werner Benger, René Heinzl, Dietmar Hildenbrand,
Tino Weinkauf, Holger Theisel, and David Tschumperlé

Index 2163

Contributors

Andy Adler Systems and Computer Engineering, Clarkson University, Ottawa, ON, Canada

Habib Ammari Department of Mathematics and Applications, Ecole Normale Supérieure, Paris, France

Mark A. Anastasio Biomedical Engineering, Illinois Institute of Technology, Chicago, IL, USA

Simon R. Arridge Department of Computer Science, University College London, London, UK

Leah Bar Department of Mathematics, Tel Aviv University, Minneapolis, MN, USA

Florian Becker Heidelberg Collaboratory for Image Processing, University of Heidelberg, Heidelberg, Germany

Werner Benger Airborne Hydromapping Software GmbH, Innsbruck, Austria

Institute for Astro- and Particle Physics, University of Innsbruck, Innsbruck, Austria

Center for Computation and Technology, Lousiana State University, Baton Rouge, LA, USA

Mario Bertero DIBRIS, Università di Genova, Genova, Italy

Liliana Borcea Department of Mathematics, University of Michigan, Ann Arbor, MI, USA

Brett Borden Department of Physics, Naval Postgraduate School of Engineering, Monterey, CA, USA

Jonathan M. Borwein School of Mathematical and Physical Sciences, University of Newcastle, Newcastle, NSW, Australia

Alexander M. Bronstein Computer Science Department, Technion-Israel Institute of Technology, Haifa, Israel

Michael M. Bronstein Computer Science Department, Technion-Israel Institute of Technology, Haifa, Israel

Antoni Buades Universitat Illes Balears, Palma de Mallorca, Spain

Martin Burger Institute for Computational and Applied Mathematics, University of Münster, Münster, Germany

Charles Byrne Department of Mathematical Sciences, University of Massachusetts Lowell, Lowell, MA, USA

Daniela Calvetti Department of Mathematics and Department of Cognitive Science, Case Western Reserve University, Cleveland, OH, USA

V. Caselles DTIC, Universitat Pompeu-Fabra, Barcelona, Spain

A. Chambolle CNRS UMR 7641, Ecole Polytechnique, Palaiseau Cedex, France

Raymond Chan Department of Mathematics, The Chinese University of Hong Kong, Shatin, Hong Kong

Tony F. Chan Office of the President, Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong

Margaret Cheney Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, NY, USA

Jin Cheng School of Mathematical Sciences, Fudan University, Shanghai, China

Ole Christensen Department of Mathematics, Technical University of Denmark, Lyngby, Denmark

Ginmo Chung Los Angeles, CA, United States

Julianne Chung Virginia Tech, Blacksburg, VA, USA

Tomeu Coll Universitat de les Illes Balears, Palma-Illes Balears, Spain

David Colton Department of Mathematical Sciences, University of Delaware, Newark, DE, USA

Daniel Cremers Department of Computer Science, Technische Universität München, Garching, Germany

Julie Digne École normale supérieure de Cachan, CMLA, Cachan, France

LIRIS, Centre National de la Recherche Scientifique (CNRS), Lyon, France

Mariella Dimiccoli Image Processing Group, Pompeu Fabra University (UPF), Barcelona, Spain

Oliver Dorn School of Mathematics, The University of Manchester, Manchester, UK

Instituto Gregorio Millán Barbany, Universidad Carlos III de Madrid, Leganés (Madrid), Spain

Paul P.B. Eggermont Food and Resource Economics, University of Delaware, Newark, DE, USA

Peter Elbau Computational Science Center, University of Vienna, Vienna, Austria

Hans G. Feichtinger University of Vienna, Vienna, Austria

Athanasios S. Fokas University of Technology Darmstadt, Darmstadt, Germany

Massimo Fornasier Faculty of Mathematics, Technische Universität München, Garching, Germany

Brigitte Forster Fakultät für Informatik und Mathematik, Universität Passau, Passau, Germany

Romina Gaburro University of Limerick, Limerick, Ireland

Charles Groetsch Traubert Chair in Science and Mathematics, The Citadel, Charleston, SC, USA

Martin Hanke-Bourgeois Institut für Mathematik, Johannes Gutenberg-Universität Mainz, Mainz, Germany

René Heinzl Shenteq s.r.o, Bratislava, Slovak Republic

Gabor T. Herman Department of Computer Science, The Graduate Center of the City University of New York, New York, NY, USA

Dietmar Hildenbrand University of Technology Darmstadt, Darmstadt, Germany

Bernd Hofmann Faculty of Mathematics, Technische Universität Chemnitz, Chemnitz, Germany

Miyoun Jung Department of Mathematics, Hankuk University of Foreign Studies, Los Angeles, CA, USA

Jari P. Kaipio Department of Mathematics, University of Auckland, Auckland, New Zealand

Barbara Kaltenbacher Institut für Mathematik, Alpen-Adria-Universität Klagenfurt, Klagenfurt, Austria

Hyeonbae Kang Department of Mathematics, Inha University, Incheon, Korea

George A. Kastis Research Center of Mathematics, Academy of Athens, Athens, Greece

Andreas Kirsch Department of Mathematics, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany

Nahum Kiryati Tel Aviv University, Tel Aviv, Israel

Sarah Knepper Emory University, Atlanta, GA, USA

Ville Kolehmainen Department of Physics and Mathematics, University of Eastern Finland, Kuopio, Finland

Rainer Kress Institut für Numerische und Angewandte Mathematik, Universität Göttingen, Göttingen, Germany

Venkateswaran P. Krishnan Centre for Applicable Mathematics, Tata Institute for Fundamental Research, Bangalore, Karnataka, India

Peter Kuchment Mathematics Department, Texas A & M University, College Station, TX, USA

Leonid Kunyansky Department of Mathematics, University of Arizona, Tucson, AZ, USA

Matti Lassas Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland

Dominique Lesselier Laboratoire des Signaux et Systemes, CNRS, Gif-sur-Yvette, France

William Lionheart The University of Manchester, Manchester, UK

D. Russell Luke Institute of Numerical and Applied Mathematics, Georg-August-Universität Göttingen, Göttingen, Germany

Leonidas Mindrinos Computational Science Center, University of Vienna, Vienna, Austria

Jan Modersitzki Institute of Mathematics and Image Computing, University of Lübeck, Luebeck, Germany

Jean-Michel Morel École Normale Supérieure de Cachan, Cachan, France

Fionn Murtagh School of Computer Science and Informatics, De Montfort University, Leicester, UK

James G. Nagy Emory University, Atlanta, GA, USA

Frank Natterer Department of Mathematics and Computer Science, University of Münster, Münster, Germany

Andreas Neubauer Industrial Mathematics Institute, Johannes Kepler University Linz, Linz, Austria

Mila Nikolova CMLA, ENS Cachan, CNRS, Cachan Cedex, France

M. Novaga Dipartimento di Matematica, Università di Padova, Padova, Italy

Ozan Öktem Department of Mathematics, KTH Royal Institute of Technology, Stockholm, Sweden

Christiane Pöschl Institute of Mathematics, Alpen Adria Universität Klagenfurt, Klagenfurt, Austria

Stephan Paukner Applied Research Center Communication Systems, GmbH, Vienna, Austria

Stefania Petra Faculty of Mathematics and Computer Science, Heidelberg University, Heidelberg, Germany

Eric Todd Quinto Department of Mathematics, Tufts University, Medford, MA, USA

Holger Rauhut Lehrstuhl C für Mathematik, RWTH Aachen University, Aachen, Germany

Martin Rumpf Institute for Numerical Simulation, Bonn University, Bonn, Germany

Lars Ruthotto Department of Mathematics and Computer Science, Emory University, Atlanta, GA, USA

Neus Sabater CMLA, École normale supérieure de Cachan, Cachan, France

Philippe Salembier Department of Signal and Communication, Universitat Politechnica de Catalunya, Barcelona, Spain

Mikko Salo Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland

Otmar Scherzer Computational Science Center, University of Vienna, Vienna, Austria

RICAM, Austrian Academy of Sciences, Linz, Austria

Christoph Schnörr Faculty of Mathematics and Computer Science, Heidelberg University, Heidelberg, Germany

Nir Sochen Tel Aviv University, Tel Aviv, Israel

Erkki Somersalo Department of Mathematics, Case Western Reserve University, Cleveland, OH, USA

Jean-Luc Starck CEA, Laboratoire AIM, CEA/DSM-CNRS-Université Paris Diderot, CEA, IRFU, Service d'Astrophysique, Centre de Saclay, Gif-Sur-Yvette Cedex, France

Gabriele Steidl Department of Mathematics, University of Kaiserslautern, Kaiserslautern, Germany

Tanja Tarvainen Department of Physics and Mathematics, University of Eastern Finland, Kuopio, Finland

Holger Theisel Institut fur Simulation und Graphik AG Visual Computing, Magdeburg, Germany

Alain Trouvé Centre de Mathématiques et Leurs Applications, Ecole Normale Superieure Cachan, Cachan Cédex, France

David Tschumperlé GREYC (UMR-CNRS 6072), CAEN Cedex, France

Gunther Uhlmann Department of Mathematics, University of Washington, Seattle, WA, USA

Luminita A. Vese Department of Mathematics, Hankuk University of Foreign Studies, Los Angeles, CA, USA

Kun Wang Medical Imaging Research Center, Illinois Institute of Technology, Chicago, IL, USA

Tino Weinkauf Feature-Based Data Analysis for Computer Graphics and Visualization, Max Planck Institute for Informatics, Saarbrücken, Germany

Benedikt Wirth Institute for Numerical Simulation, Bonn University, Bonn, Germany

Andy Yip Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong

Laurent Younes Department of Applied Mathematics and Statistics, The Johns Hopkins University, Baltimore, MD, USA

Part I

Inverse Problems – Methods

Linear Inverse Problems

Charles Groetsch

Contents

1 Introduction 4
2 Background 6
3 Mathematical Modeling and Analysis 11
A Platonic Inverse Problem 11
Cormack's Inverse Problem 14
Forward and Reverse Diffusion 16
Deblurring as an Inverse Problem 17
Extrapolation of Band-Limited Signals 19
PET 20
Some Mathematics for Inverse Problems 21
4 Numerical Methods 32
Tikhonov Regularization 32
Iterative Regularization 37
Discretization 39
5 Conclusion 43
Cross-References 44
References 44

Abstract

This introductory treatment of linear inverse problems is aimed at students and neophytes. A historical survey of inverse problems and some examples of model inverse problems related to imaging are discussed to furnish context and texture to the mathematical theory that follows. The development takes place within the sphere of the theory of compact linear operators on Hilbert space, and the singular value decomposition plays an essential role. The primary concern is regularization theory: the construction of convergent well-posed approximations

C. Groetsch (✉)
Traubert Chair in Science and Mathematics, The Citadel, Charleston, SC, USA
e-mail: charles.groetsch@citadel.edu

© Springer Science+Business Media New York 2015

O. Scherzer (ed.), *Handbook of Mathematical Methods in Imaging*,
DOI 10.1007/978-1-4939-0790-8_1

to ill-posed problems. For the most part, the discussion is limited to the familiar regularization method devised by Tikhonov and Phillips.

1 Introduction

> ► . . . although nature begins with the cause and ends with the experience we must follow the opposite course, namely . . . begin with the experience and by means of it end with the cause.
> Leonardo da Vinci

An inverse problem is the flip side of some direct problem. Direct problems treat the transformation of known causes into effects that are determined by some specified model of a natural process. They tend to be future directed and outward looking and are often concerned with forecasting or with determining external effects of internal causes. Direct problems have solutions (causes have effects), and the process of transforming causes into effects is a mathematical *function*: a given cause determines, via the model, a unique effect. In direct problems the operator that maps causes into effects is typically continuous in natural metrics: close causes have close effects. These features of direct problems make them *well posed*.

The idea of a well-posed problem has its origins in Jacques Hadamard's short paper [37] published in 1902. Hadamard held the opinion that an important physical problem must have three attributes:

1. (Existence) It has a solution.
2. (Uniqueness) It has only one solution.
3. (Stability) The solution depends continuously on the data of the problem.

A problem satisfying these three conditions is called *well posed*. In his 1902 paper, Hadamard called a problem *bien posé* if it has properties (1) and (2). Again in his 1923 lectures [38], he called a problem "correctly set" if it satisfies (1) and (2). Condition (3) was not named as a specific requirement of a well-posed problem, but his explicit notice of the lack of continuous dependence on boundary data of the solution of Cauchy's problem for Laplace's equation led to (3) becoming part of the accepted definition of a well-posed problem.

A problem is *ill posed* if it lacks these qualities. Hadamard's suggestion that ill-posed problems are devoid of physical significance (*déprourvu de signification physique*) was unfortunate, as almost all inverse problems in the physical and biological sciences are ill posed. To be fair, it should be noted that Hadamard was speaking about a specific problem, the Cauchy problem for Laplace's equation in a strip. On the other hand, Courant [15] insisted more generally that "a mathematical problem cannot be considered as realistically corresponding to physical phenomena

unless . . . " it satisfies condition (3). The problems of existence and uniqueness in inverse problems can often be ameliorated by generalizing the notion of solution and constraining the generalized solution, but the key attribute of stability often is a feature that is inherently absent in inverse problems. This essential lack of stability usually has dire consequences when numerical methods, using measured or uncertain data, are applied to inverse problems.

Inverse problems are as old as science itself. In fact, a reasonable working definition of science is the explanation of natural phenomena by the construction of conceptual models for interpreting imperfect observational representations of "true" natural objects or processes. This definition encompasses the three essential ingredients of mathematical inverse problems: a "true" solution, a model, or operator that transforms this true solution into an imperfect representation that is amenable to observations or measurements. One could say that inverse theory embraces an operating principle that is essentially Platonic: true natural objects exist, but it is only through models and imperfectly perceived images that we experience them. The challenge is to "invert" the model to recover a useful estimate of the true object from the observed image. In this sense, all of inverse theory deals with "imaging."

A mathematical framework for the study of inverse problems must provide sufficient scope for each of the three elements: true solutions, model, and observations. In this chapter the solution space and the space of observations are both taken to be Hilbert spaces, but not necessarily the same Hilbert space, as one naturally desires more of the solution than one demands from the observations. The model is a transformation or operator that carries a possible solution to an observed effect. We consider only linear inverse problems, so our models are linear operators.

Any practical model suppresses some information. If a model represents every bit of information in the objects themselves (i.e., the model operator is the identity operator), then nothing is gained in conceptual economy. In this case one is in the absurd position of Mein Herr in Lewis Carroll's *Sylvie and Bruno Concluded*:

> "We actually made a map of the country, on a scale of a *mile to the mile*!" . . . "It has never been spread out yet," said Mein Herr: "the farmers objected; they said it would cover the whole country, and shut out the sunlight! So now we use the country itself, as its own map, and I assure you it does nearly as well."

Finite linear models lead to linear algebra problems. Idealized limiting versions of finite models typically lead to compact linear operators, that is, limits of finite rank operators. A compact operator may have a nontrivial null-space, a non-closed range, or an unbounded (generalized) inverse. Therefore, these operators, which occur widely in models of linear inverse problems, lack all the virtues of well posedness. In this chapter, we provide a somewhat slanted survey of linear inverse problems, mainly involving compact operators, with special attention to concepts underlying methods for constructing stable approximate solutions.

Before draping these ideas on a mathematical framework, we discuss a half dozen examples of model inverse problems that have played significant roles in the development of the physical sciences.

2 Background

> ► Our science is from the watching of shadows;
> Ezra Pound

This brief and incomplete historical survey of physical inverse problems is meant to give some perspective on certain inverse problems closely related to imaging in the broad sense. Our viewpoint involves both the very large scale, treating inverse problems loosely associated with assessing the cosmos, and the human scale, dealing with evaluation of the inaccessible interior of bodies (human or otherwise).

Inverse theory, as a distinct field of inquiry, is a relatively recent development; however, inverse problems are as old as science itself. A desire to know causes of perceived effects is ingrained in the human intellectual makeup. The earliest attempts at explanations, as, for example, in the creation myths of various cultures, were supernatural – grounded in mysticism and mythology. When humankind embarked on a program of rationalization of natural phenomena, inverse problems emerged naturally and inevitably. An early example is Plato's allegory of the cave (ca. 375 B.C.). In the seventh book of his *Republic*, Plato describes the situation. A group of people have been imprisoned since their birth in a cave where they are chained in such a manner that allows them to view only a wall at the back of the cave. Outside the cave life goes on, illuminated by a fire blazing in the distance. The captives in the cave must reconstruct this external reality on the basis of shadows cast on the rear wall of the cave. This is the classic inverse imaging problem: real objects are perceived only as two-dimensional images in the form of shadows on the cave wall. This annihilation of a dimension immediately implies that the reconstruction problem has multiple solutions and that solutions are unstable in that highly disparate objects may have virtually identical images.

Aristotle adapted his teacher Plato's story of the cave to address a scientific inverse problem: the shape of the earth. This shape could not be *directly* assessed in Aristotle's time, so he suggested an *indirect* approach (see Book II of *On the Heavens*.):

> As it is, the shapes which the moon itself each month shows are of every kind – straight, gibbous, and concave – but in eclipses the outline is always curved; and since it is the interposition of the earth that makes the eclipse, the form of the line will be caused by the form of the earth's surface, which is therefore spherical.

Aristotle's reasoning provided an *indirect* argument for the sphericity of the earth based on the shapes of shadows cast on the moon.

Inverse imaging has been a technical challenge for centuries. The difficulties that early investigators encountered were vividly captured by Albrecht Dürer's woodcut *Man Drawing a Lute* (1525). We can see the doubts and angst brought on by the inverse imaging problem etched on the face of the crouching technician (Fig. 1).

The character on the left, standing bolt upright in a confident attitude, has the comparatively easy direct problem. He has complete knowledge of the object, and he knows exactly how the projection model will produce the image. On the other hand, the crouching man on the right, with the furrowed brow, faces the more difficult

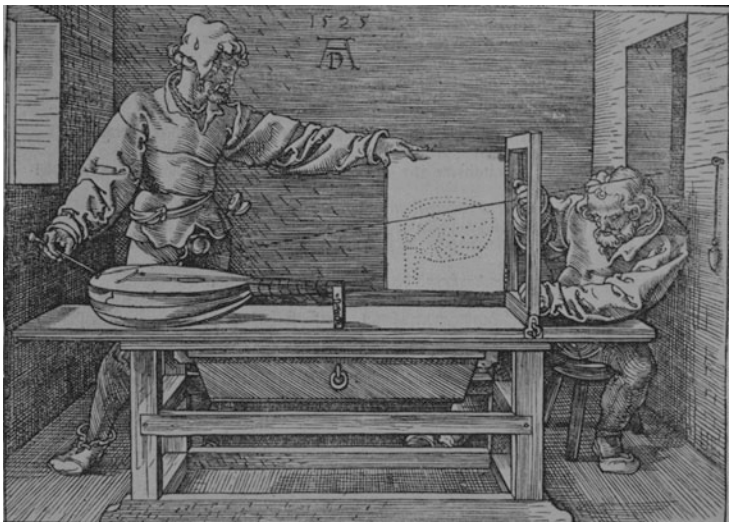


Fig. 1 A renaissance inverse problem

inverse problem of assessing whether the image captures the essential features of the object. Dürer's woodcut is a striking representation of the comparative difficulties of direct and inverse assessment.

Modern imaging science has its roots in Galileo Galilei's lunar observations carried out during the winter of 1609. Prior to Galileo's study, the moon was thought to belong to the realm of the Pythagorean fifth essence, consisting of perfectly uniform material in perfect spherical form. Galileo's eyes, empowered by his improved telescope, the earliest scientific imaging device, showed him otherwise [21]:

> . . . we certainly see the surface of the Moon to be not smooth, even, and perfectly spherical, as the great crowd of philosophers has believed about this and other heavenly bodies, but, on the contrary, to be uneven, rough, and crowded with depressions and bulges. And it is like the Earth itself, which is marked here and there with chains of mountains and depth of valleys.

But Galileo was not satisfied with qualitative evidence. He famously stated that the book of nature is written in the language of mathematics, and he used mathematics, in the form of the Pythagorean theorem, along with some shrewd estimates, to assess indirectly the heights of lunar mountains. The process of indirect assessment is a hallmark of inverse problems in the natural sciences. (See [2] for an account of inverse problems of indirect assessment.)

Nonuniqueness is a feature of many inverse problems that was slow to gain acceptance. An early instance of this phenomenon in a physical inverse problem occurred in the kinematic studies of ballistics carried out by Niccolò Tartaglia in the sixteenth century. Tartaglia claimed to be the inventor of the gunner's square, a device for measuring the angle of inclination of a cannon. Using his square Tartaglia

carried out ranging trials and published some of the earliest firing tables. He studied not only the direct problem of finding ranges for a given firing angle but also the inverse problem of determining the firing angle that results in a given range. Although Tartaglia's treatment was conceptually flawed and lacked rigor, his work contains glimmers of a number of basic principles of mathematical analysis that took several centuries to mature [32]. Tartaglia was particularly struck by nonuniqueness of solutions of the inverse problem. As he put it proudly in the dedication of his book *Nova Scientia* (Venice, 1537):

> I knew that a cannon could strike in the same place with two different elevations or aimings, I found a way of bringing about this event, a thing not heard of and not thought by any other, ancient or modern.

With this boast Tartaglia was one of the first to call attention to this common feature of nonuniqueness in inverse problems.

Tartaglia found that for a given fixed charge, each range (other than the maximum range) is achieved by two distinct aimings placed symmetrically above and below the 45° inclination. A century and a half later, Edmond Halley [39] took up the more general problem of allowing both the charge and the firing angle to vary while firing on a fixed target situated on an inclined plane. In this case the inverse problem of determining charge-angle pairs that result in a strike on the target has infinitely many solutions. (Of course, Halley did not address air resistance; his results are extended to the case of first-order resistance in [33].) Halley restored uniqueness to the inverse aiming problem by restricting consideration to the solution which minimizes what we would now call the kinetic energy of the emergent cannon ball. The idea of producing uniqueness by seeking the solution that minimizes a quadratic functional would in due course become a key feature of inverse theory.

The model for a modern scientific society was laid out in Francis Bacon's utopian novel *The New Atlantis* (1626). Bacon describes a voyage to the mythical land of Bensalem, which was inhabited by wise men inclined to inverse thinking. Solomon's House, a research institute in Bensalem, was dedicated to the "knowledge of *causes*, and secret motions of things; and the enlarging of the bounds of human empire, to the *effecting* of all things possible" (my italics). The Royal Society of London, founded in 1660 and modeled on Baconian principles, was similarly dedicated. The first great triumph (due largely to Halley's efforts) of the Royal Society was the publication of Newton's magisterial *Principia Mathematica* (1687). In the *Principia*, Newton's laws relating force, mass, and acceleration, combined with his inverse square law of gravity, were marshaled to solve the direct problem of two-body dynamics, confirming the curious form of Kepler's planetary orbits: an inverse square centrally directed force leads to an orbit, which is a conic section. But Newton was not satisfied with this. He also treated the inverse problem of determining what gravitational law (cause) can give rise to a given geometrical orbit (effect).

In the history of science literature, the problem of determining the orbit, given the law of attraction, is sometimes called the inverse problem; this practice inverts the terminology currently common in the scientific community. The reverse terminology in the history community is evidently a consequence of the fact that

Newton took up the determination of force law first and then treated the orbit determination problem. Indeed, Newton treated the inverse problem of orbits before he took up the direct problem. After all, his primary goal was to discover the laws of nature, the causes, rather than the effects. As Newton put it in the preface to the first edition of his *Principia*: " . . . the whole burden of philosophy seems to consist of this – from the phenomena of motions to investigate the forces of nature, and then from these forces to demonstrate the other phenomena."

In 1846, mathematical inverse theory produced a spectacular scientific triumph – the discovery of another world. The seeds of the discovery lay in the observed irregularities in the orbit of Uranus, the most distant of the planets known at the time. The orbit of Uranus did not fit with predictions based on Newton's theories of gravity and motion. In particular an orbit calculated to fit contemporary observations did not fit observations made in the previous century, and an orbit that fit to the older sightings did not match the contemporary data. This suggested two possibilities: either Newton's theory had to be modified at great distances or perhaps the anomalies in the orbit of Uranus were the effect of an as yet undiscovered planet (the cause) operating on Uranus via Newton's laws.

During the summer vacation of 1841, John Couch Adams, an undergraduate of St. John's College, Cambridge, was intrigued by the second possibility. He recorded this diary entry:

> 1841, July 3. Formed a design in the beginning of the week, of investigating, as soon as possible after taking my degree, the irregularities in the motion of Uranus, which are yet unaccounted for; in order to find whether they may be attributed to the action of an undiscovered planet beyond it; and if possible thence to determine the elements of its orbit, etc. approximately, which would probably lead to its discovery.

Adams solved the inverse problem of determining the characteristics of the orbit of the undiscovered planet, now known as Neptune, that perturbs Uranus. However, a sequence of lamentable missteps, involving his own timidity, bureaucratic inertia, and other human factors, resulted in the honor of "discovering" the new planet on the basis of mathematics going to Urbain Le Verrier of France, who solved the inverse problem independently of Adams. This of course led to disappointment in England over the botched opportunity to claim the discovery and to a good deal of hauteur in France over the perceived attempt by the English to grab credit deserved by a Frenchman. The fascinating story of the unseemly squabble is well told in [36]. See also [63] for a recent update in which old wounds are reopened.

Newton's discussion of inverse orbit problems in his *Principia*, and vague doubts about the form of the gravitational force law raised prior to the discovery of Neptune, may have inspired other inverse problems. An early interesting "toy" inverse problem in this vein was published by Ferdinand Joachimstahl in 1861 [47]. The problem Joachimstahl posed, and solved by an Abel transform, was to determine the law of gravitational attraction if the total force at any distance from a line of known mass density is given.

Johann Radon laid the foundation of mathematical imaging science, without knowing it, in his 1917 memoir [61]. (An English translation of Radon's paper may be found in [17].) Radon was concerned with the purely mathematical problem of

determining a real-valued function of two variables from knowledge of the values of its line integrals over all lines intersecting its domain. Although Radon evidently had no application in mind, his treatment was to become, after its rediscovery a half century later, the basis for the mathematics of computed tomography. (See [14] for more on the history of computed tomography.) Essentially the same result was obtained independently by Viktor Ambartsumian [1] who was interested in a specific inverse problem in astronomy. Proper motions of stars are difficult to determine, but radial velocities (relative to the earth) are obtainable from chromatic Doppler shift measurements. Ambartsumian used a mathematical model essentially equivalent to that of Radon to deduce the true three-dimensional distribution of stellar velocities from the distribution of the radial velocities.

In the mid-1950s of the last century, Allan Cormack, a young physics lecturer at the University of Cape Town, who was moonlighting in the radiology department of Groote Schuur Hospital, had a bright idea. In Cormack's words:

> It occurred to me that in order to improve treatment planning one had to know the distribution of the attenuation coefficient of tissues in the body, and that this distribution had to be found by measurements made external to the body. It soon occurred to me that this information would be useful for diagnostic purposes and would constitute a tomogram or series of tomograms, though I did not learn the word "tomogram" for many years.

This was the birth of the mathematical theory of medical imaging. Cormack would not learn of Radon's work for another two decades, but he developed the basic results for radially symmetric attenuation coefficient distributions and tested the theory with good results on a simple manufactured specimen in the form of a cylinder of aluminum encased in an annular prism of wood. The reconstructed piecewise constant attenuation function matched that of the known specimen well enough to show the promise of this revolutionary new imaging technology.

In the 1990s, inverse thinking and indirect assessment led to another spectacular advance in astronomy: the discovery of extrasolar planets. Philosophers had speculated on the reality of planets linked to the stars at least since classical Greek times, and few in modern times doubted the existence of extrasolar planets. But convincing evidence of their existence had to await the development of sufficiently sensitive telescope-mounted spectrometers and the application of simple inverse theory. The indirect evidence of extrasolar planets consisted of spectral shift data extracted from optical observations of a star.

In a single star–planet system, determining the variable radial velocity (relative to the earth) of a star wobbling under the gravitational influence of an orbiting planet of known mass and orbital radius is a simple direct problem – just equate the gravitational acceleration of the planet to its centripetal acceleration. (Consider only the simple case in which the planet, star, and earth are coplanar and the orbit is circular; an orbit oblique to the line of sight from earth introduces an additional unknown quantity. As a consequence of this obliquity, the relative mass estimated from the inverse problem is actually a *lower* bound for this quantity.) Using Doppler shift data, a simple inverse problem model may be developed for determining approximations to the relative planetary mass and orbital radius. The solution of the inverse problem enabled astronomers to announce in 1995 the existence of the first

confirmed extrasolar planet orbiting the star 51Pegasi. The millennia-old question of the existence of extrasolar worlds finally had a convincing positive answer.

We bring this historical survey of inverse problems up to the present day with the greatest challenge in contemporary cosmology: the search for dark matter. Such matter, being "dark," is by definition inaccessible to direct measurement. But recently an imaging model on the largest scale in the history of science has come to be used in attempts to assay this dark matter. The process of gravitational lensing, which is based on Einstein's theory of curved space-time, presents the possibility of inverting the imaging model to estimate a dark mass (the gravitational lens) that intervenes between the observer on earth and an immensely distant light source. The dark mass warps space in its vicinity resulting, under appropriate conditions, in focusing onto the earth light rays that in flat space would not intersect the earth. In an extreme case in which the light source (e.g., a galaxy), the intervening gravitational lens (dark matter), and the collected image are collinear, this results in a phenomenon called an Einstein ring (first observed in 1979; see [22]). If the distances from earth to the lens and from the lens to source can be estimated, then the solution of an inverse problem gives an estimate of the dark mass (see [56]).

3 Mathematical Modeling and Analysis

> ► . . . we have to remember that what we observe is not nature in itself but nature exposed to our method of questioning.
> Werner Heisenberg

A Platonic Inverse Problem

Plato's discussion of captives struggling to discern the real cause of shadows cast on the back wall of a cave is a primeval exemplar of inverse imaging problems. Here we present a toy imaging problem inspired by Plato's allegory. While the problem is very elementary, it usefully illustrates some important aspects of imaging problems and inverse problems in general.

Imagine a two-dimensional convex object in the xy-plane, which is bounded by the positive coordinate axes, and the graph of a function $y = f(x), 0 \leq x \leq 1$ that is positive on $[0, 1)$, strictly decreasing and concave-down, and satisfies $f(1) = 0 = f'(0)$. The object is illuminated by parallel light rays from the left that form angles θ with the negative ray of the horizontal axis, as illustrated in Fig. 2.

The goal is to reconstruct the shape of the object from observations of the extent $s(\theta)$ of the shadow cast by the object. This is accomplished by fashioning a parameterization $(x(\theta), f(x(\theta)))$ of the boundary curve of the object. As a further simplification we assume that $f'(1) = -1$. These assumptions guarantee that for each $s > 1$ there is a unique point $(t, f(t))$ on the graph of f at which the tangent line intersects the x-axis at s. What is required to see this is the existence of a unique

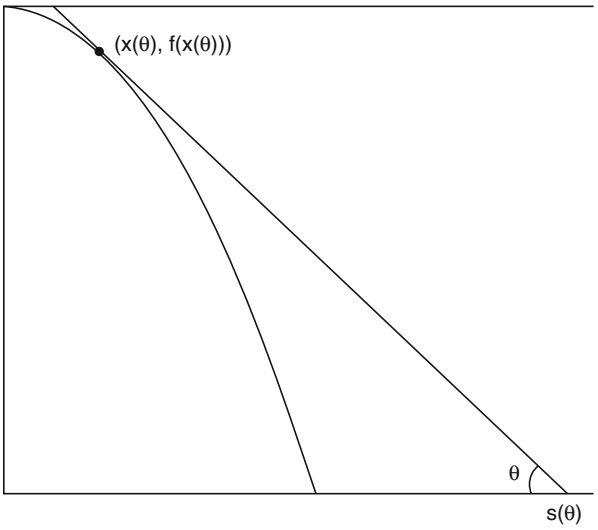


Fig. 2 A model shadow problem

$t \in (0, 1)$ such that the tangent line to the graph at the point $(t, f(t))$ intersects the x-axis at s. That is,

$$(s - t) f'(t) + f(t) = 0.$$

For each fixed $s > 1$, the expression on the left is strictly decreasing for $t \in (0, 1)$, positive at $t = 0$ and negative at $t = 1$, so the existence of a unique such $t = x(\theta)$ is assured. At the point of tangency,

$$-\tan\theta = f'(x(\theta)).$$

Also,

$$f(x(\theta)) = (\tan\theta)(s(\theta) - x(\theta)),$$

and hence determining $x(\theta)$ also gives $(x(\theta), f(x(\theta)))$, which solves the inverse imaging problem. Combining these results we have

$$-(\tan\theta)x'(\theta) = f'(x(\theta))x'(\theta) = (s(\theta) - x(\theta))\sec^2\theta + (s'(\theta) - x'(\theta))\tan\theta.$$

A bit of simplification yields

$$x(\theta) = s(\theta) + \frac{1}{2}\sin(2\theta)s'(\theta), \tag{1}$$

which explicitly solves the inverse problem of determining the shape $(x(\theta), f(x(\theta)))$ from knowledge of the extent of the shadows $s(\theta)$.

The explicit formula (1) would seem to completely solve the inverse problem. In a theoretical sense this is certainly true. However, the formulation (1) shelters a subversive factor (the derivative) that should alert us to potential challenges involved in the practical solution of the inverse problem. Observations are always subject to measurement errors. The differentiation process, even if performed exactly, may amplify these errors as differentiation is a notoriously unstable process. For example, if a shadow function $s(\theta)$ is perturbed by low-amplitude high-frequency noise of the form $\eta_n(\theta) = \frac{1}{n} \sin n^2\theta$ giving observed data

$$s_n(\theta) = s(\theta) + \eta_n(\theta),$$

then the corresponding shape abscissas provided by (1) satisfy

$$x_n(\theta) = x(\theta) + \eta_n(\theta) + \frac{\sin 2\theta}{2}\eta_n'(\theta).$$

But η_n converges uniformly to 0 as $n \to \infty$, while $\max|\eta_n'| \to \infty$, giving a convincing illustration of the instability of the solution of the inverse problem provided by (1). For more examples of model inverse problems with explicit solutions that involve differentiation, see [71].

It is instructive to view the inverse problem from another perspective. Note that by (1), $s(\theta)$ is the solution of the linear differential equation

$$\frac{ds}{d\theta} + \frac{2}{\sin(2\theta)}s = \frac{2}{\sin(2\theta)}x(\theta)$$

satisfying $s(\pi/4) = 1$. This differential equation may be solved by elementary means yielding

$$s(\theta) = \frac{1 + \cos 2\theta}{\sin 2\theta} + \int_{\pi/4}^{\theta} \frac{2(1 + \cos 2\theta)}{(1 + \cos 2\varphi)\sin 2\theta} x(\varphi)\, d\varphi. \tag{2}$$

In this formulation, the "hidden" solution $x(\varphi)$ of the inverse problem is seen to be transformed by a linear integral operator into observations of the shadows $s(\theta)$. The goal now is to uncover $x(\varphi)$ from (2) using knowledge of $s(\theta)$, that is, one must solve an integral equation.

The solution of the integral formulation (2) suffers from the same instability as the explicit solution (1). Indeed, one may write (2) as

$$s = \psi + \psi T x$$

where $\psi(\theta) = (1 + \cos 2\theta)/\sin 2\theta$ and

$$(Tx)(\theta) = \int_{\pi/4}^{\theta} \frac{2}{1 + \cos 2\varphi} x(\varphi)\, d\varphi.$$

If we let $v_n(\varphi) = \frac{n}{2}(1 + \cos 2\varphi) \sin n^2\varphi$ and set

$$s_n = \psi + \psi T(x + v_n)$$

then one finds that $s_n \to s$ uniformly, while $\max |v_n| \to \infty$. That is, arbitrarily small perturbations in the data s may correspond to arbitrarily large deviations in the solution x. This story has a moral: instability is intrinsic to the inverse problem itself and not a manifestation of a particular representation of the solution.

Cormack's Inverse Problem

As noted in the previous section, the earliest tomographic test problem explicitly motivated by medical imaging was Cormack's experiment [12] with a fabricated sample having a simple radially symmetric absorption coefficient. The absorption coefficient is a scalar field whose support may be assumed to be contained within the body to be imaged. This coefficient f supplies a measure of the attenuation rate of radiation as it passes through a given body point and is characterized by Bouguer's law

$$\frac{dI}{ds} = -fI,$$

where I is the intensity of the radiation and s is arclength. The integral of f along a line L intersecting the body then satisfies

$$g = \int_L f\, ds.$$

Here $g = \ln(I_0/I_e)$, where I_0 is the incident intensity, and I_e the emergent intensity, of the beam. The observable quantity g is then a measure of the total attenuation effect that the body has on the beam traversing the line L.

To be more specific, for a given $t \in \mathbf{R}$ and a given unit vector $\vec{\varphi} = (\cos\varphi, \sin\varphi)$, let $L_{t,\varphi}$ represent the line

$$L_{t,\varphi} = \{\vec{x} \in \mathbf{R}^2 : \langle \vec{x}, \vec{\varphi} \rangle = t\}$$

where $\langle \cdot, \cdot \rangle$ is the Euclidean inner product. We will denote the integral of f over $L_{t,\varphi}$ by

$$\mathcal{R}(f)(t,\varphi) = \int_{L_{t,\varphi}} f\, ds = \int_{-\infty}^{\infty} f(t\cos\varphi - s\sin\varphi, t\sin\varphi + s\cos\varphi)\, ds.$$

If f is *radial*, that is, independent of φ, then

$$\mathcal{R}(f)(t,\varphi) = \mathcal{R}(f)(t,0) = \int_{-\infty}^{\infty} f(t,s)\, ds. \tag{3}$$

Furthermore, if f vanishes exterior to the disk of radius R, then on setting $r = \sqrt{t^2+s^2}$ and $f(r) = f(t,s)$, one finds

$$g(t) = \int_t^R \frac{2rf(r)}{\sqrt{r^2-t^2}}\, dr, \tag{4}$$

where $g(t) = \mathcal{R}(f)(t,0)$. The mapping defined by (4), which for a given line $L_{t,\varphi}$ transforms the radial attenuation coefficient into the function g, is an *Abel transform* of f. It represents, as a direct problem, Cormack's early experiment with a radially symmetric test body. Determining the distribution of the attenuation coefficient requires solving the inverse problem. The Abel transform may be formally inverted by elementary means to furnish a solution of the inverse problem of determining the attenuation coefficient f from knowledge of the loss data g. Indeed, by (4) and a reversal of order of integration,

$$\int_r^R \frac{tg(t)}{\sqrt{t^2-r^2}}\, dt = \int_r^R f(s)s \int_r^s \frac{2t}{\sqrt{s^2-t^2}\sqrt{t^2-r^2}}\, dt\, ds = \pi \int_r^R f(s)s\, ds,$$

since

$$\int_r^s \frac{2t}{\sqrt{s^2-t^2}\sqrt{t^2-r^2}}\, dt = \pi$$

(change the variable of integration to $w = \sqrt{s^2-t^2}/\sqrt{s^2-r^2}$). However,

$$\int_r^R \frac{tg(t)}{\sqrt{t^2-r^2}}\, dt = -\int_r^R (t^2-r^2)^{1/2} g'(t)\, dt$$

and hence on differentiating, we have

$$\int_r^R \frac{rg'(t)}{\sqrt{t^2-r^2}} dt = -\pi r f(r).$$

Therefore,

$$f(r) = -\frac{1}{\pi}\int_r^R \frac{g'(t)}{\sqrt{t^2-r^2}}\, dt.$$

The derivative lurking within this inversion formula is again a harbinger of instability in the solution of the inverse problem.

Forward and Reverse Diffusion

Imagine a bar, identified with the interval $[0, \pi]$ of the x-axis, the lateral surface of which is thermally insulated while its ends are kept always at temperature zero. The diffusion of heat in the bar is governed by the one-dimensional heat equation

$$\frac{\partial u}{\partial t} = \kappa \frac{\partial^2 u}{\partial x^2}, \quad 0 < x < \pi$$

where $u(x, t)$ is the temperature at position x and time t and κ is the thermal diffusivity. If the initial temperature distribution in the bar is a function $f(x)$, then the boundary and initial conditions associated with this model are

$$u(0, t) = 0, \quad u(\pi, t) = 0, \quad u(x, 0) = f(x).$$

In the forward diffusion problem, the goal is to find, for a given future time $T > 0$, the temperature distribution $g(x) = u(x, T)$. Formal separation of variable techniques leads to a solution of the form

$$u(x, t) = \sum_{n=1}^{\infty} a_n e^{-\kappa n^2 t} \sin nx,$$

where a_n are the Fourier coefficients of the initial temperature distribution

$$a_n = \frac{2}{\pi} \int_0^{\pi} f(s) \sin ns \, ds.$$

The future temperature distribution is then seen to be, after some rearranging,

$$g(x) = \int_0^{\pi} k(x, s) f(s) \, ds,$$

where

$$k(x, s) = \frac{2}{\pi} \sum_{n=1}^{\infty} e^{-\kappa n^2 T} \sin nx \sin ns.$$

A high degree of smoothing is a notable feature of the forward diffusion process. Specifically, the factors $e^{-\kappa n^2 T}$ in the transformation have the effect of severely damping high-frequency components in the initial temperature distribution f.

A corresponding reverse diffusion process is immediately suggested, namely, the retrodiction of the initial temperature distribution f, from knowledge of the later temperature distribution g. In this inverse problem, one finds that

$$f(x) = \frac{2}{\pi} \sum_{n=1}^{\infty} e^{\kappa n^2 T} \int_0^{\pi} g(s) \sin ns \, ds. \tag{5}$$

The contrast with the forward problem is striking: now high-frequency components in g are *amplified* by the huge factors $e^{\kappa n^2 T}$. Also note that the inverse problem is soluble only for a restricted class of functions g – those for which the series (5) converges in $L^2[0, \pi]$. As will be seen in the next section, the reverse diffusion process is a useful metaphor in the discussion of deblurring.

Deblurring as an Inverse Problem

Cameras and other optical imagers capture a scene, or *object*, and convert it into an imperfect *image*. The object may be represented mathematically by a function $f : \mathbf{R}^2 \to \mathbf{R}$ that codes, for example, gray scale or intensity. The image produced by the device is a function $g : \mathbf{R}^2 \to \mathbf{R}$, and the process may be phrased abstractly as $g = Kf$, where K is an operator modeling the operation of the imager. In a perfect imager, $K = I$, the identity operator (recall Mein Herr's map!). The perfect model may be expressed in terms of the two-dimensional delta distribution as

$$f(\vec{x}) = \int \int_{\mathbf{R}^2} \delta(\vec{x} - \vec{\xi}) f(\xi) \, d\xi.$$

However, any physical imaging device blurs the object f into an image g, which in many cases can be represented by

$$g(\vec{x}) = \int \int_{\mathbf{R}^2} k(\vec{x} - \vec{\xi}) f(\xi) d\vec{\xi} \tag{6}$$

where $k(\cdot)$, the *point spread function* of the device, is some approximation of the delta function centered at the origin. Theoretical examples of such approximations include the tin-can function $\chi_R/(\pi R^2)$, where χ_R is the indicator function of the disk of radius R centered at the origin, and the sinc and sombrero functions given in polar coordinates by

$$\text{sinc}(r, \theta) = \frac{\sin \pi r}{\pi r} \quad \text{and} \quad \text{somb}(r, \theta) = 2 \frac{J_1(\pi r)}{\pi r},$$

respectively, where J_1 is the Bessel function of first kind and order 1. A frequently occurring model uses the Gaussian point spread function

$$k(r,\theta) = \frac{1}{2\pi\sigma^2} e^{-r^2/2\sigma^2}.$$

The problem of deblurring consists of solving (6) for the object f, given the blurred image g. For a good introduction to deblurring, see [44].

Reverse diffusion in two dimensions is a close cousin of deblurring. A basic tool in the analysis is the 2D Fourier transform defined for $f : \mathbf{R}^2 \to \mathbf{R}$ and $\vec{x}, \vec{\omega} \in \mathbf{R}^2$ by

$$\widehat{f}(\vec{\omega}) = \mathcal{F}\{f\}(\vec{\omega}) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-i\langle \vec{x},\vec{\omega}\rangle} f(\vec{x})\, dx_1\, dx_2$$

with the inversion formula

$$f(\vec{x}) = \frac{1}{(2\pi)^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{i\langle \vec{x},\vec{\omega}\rangle} \widehat{f}(\vec{\omega})\, d\omega_1\, d\omega_2.$$

On integrating by parts, one sees that

$$\mathcal{F}\{\Delta f\}(\vec{\omega}) = -\|\vec{\omega}\|^2 \widehat{f}(\vec{\omega}),$$

where Δ is the Laplacian operator: $\Delta = \frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial x_2^2}$. Consider now the initial value problem for the 2D heat equation

$$\frac{\partial u}{\partial t} = \kappa \Delta u \quad \vec{x} \in \mathbf{R}^2, \quad t > 0, \quad u(\vec{x}, 0) = f(\vec{x}).$$

Applying the Fourier transform yields the initial value problem

$$\frac{dU}{dt} = -\kappa\|\vec{\omega}\|^2 U, \quad U(0) = \widehat{f}$$

where $U(t) = \widehat{u}(\cdot, t)$ and hence $U(t) = \widehat{f} e^{-\|\omega\|^2\kappa t}$. The convolution theorem then gives

$$u(\vec{x},t) = \mathcal{F}^{-1}\{e^{-\|\omega\|^2\kappa t}\widehat{f}\} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} k(\vec{x}-\vec{\xi}) f(\vec{\xi})\, d\xi_1\, d\xi_2$$

where (using the integral result of [25], 12.A)

$$k(\vec{x}) = \mathcal{F}^{-1}\left\{e^{-\omega_1^2\kappa t} e^{-\omega_2^2\kappa t}\right\} = \frac{1}{4\pi^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{i\langle \vec{x},\vec{\omega}\rangle} e^{-\left(\omega_1^2+\omega_2^2\right)\kappa t}\, d\omega_1\, d\omega_2$$

$$= \frac{1}{4\pi^2}\int_{-\infty}^{\infty} e^{ix_1\omega_1 - \omega_1^2\kappa t} d\omega_1 \int_{-\infty}^{\infty} e^{ix_2\omega_2 - \omega_2^2\kappa t} d\omega_2 = \frac{1}{4\pi\kappa t} e^{-\left(x_1^2+x_2^2\right)/(4\kappa t)}.$$

The inverse problem of determining the initial distribution $f(\vec{x}) = u(\vec{x}, 0)$, given the distribution $g(\vec{x}) = u(\vec{x}, T)$ at a later time $T > 0$, is equivalent to solving the integral equation of the first kind

$$g(\vec{x}) = \frac{1}{4\pi\kappa T} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-((x_1-\xi_1)^2+(x_2-\xi_2)^2)/(4\kappa T)} f(\xi_1, \xi_2)\, d\xi_1\, d\xi_2,$$

which is in turn equivalent to the deblurring problem with Gaussian point spread function

$$\Gamma_\sigma(\vec{x}) = \frac{1}{2\pi\sigma^2} e^{-\|\vec{x}\|^2/2\sigma^2}$$

having variance $\sigma^2 = 2\kappa T$. The idea of deblurring by reverse diffusion is developed in [8].

Extrapolation of Band-Limited Signals

Extrapolation is a basic challenge in signal analysis. The Fourier transform, $\mathcal{F}$, is the analytical workhorse in this field. It transforms a time signal $f(t), -\infty < t < \infty$, into a complex-frequency distribution $\widehat{f}(\omega)$ via the formula

$$\widehat{f}(\omega) = \mathcal{F}\{f\}(\omega) = \int_{-\infty}^{\infty} f(t) e^{-i\omega t}\, dt.$$

In a suitable setting, the time-to-frequency transformation may be inverted by the formula (e.g., [25])

$$f(t) = \mathcal{F}^{-1}\{\widehat{f}\}(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \widehat{f}(\omega) e^{i\omega t}\, dt.$$

Any physically realizable detector is capable of picking up frequencies only in a limited range, say $|\omega| \le \Omega$. A signal f whose Fourier transform vanishes for $|\omega| > \Omega$, for some given $\Omega > 0$, is called a *band-limited* signal. A detector that operates in the frequency band $-\Omega \le \omega \le \Omega$ band-limits signals it collects, that is, it treats only $\chi_{[-\Omega,\Omega]}\widehat{f}$, where

$$\chi_{[-\Omega,\Omega]}(\omega) = \begin{cases} 1, & \omega \in [-\Omega, \Omega] \\ 0, & \omega \notin [-\Omega, \Omega]. \end{cases}$$

is the indicator function of the interval $[-\Omega, \Omega]$. Multiplication by $\chi_{[-\Omega,\Omega]}$ in the frequency domain is called a low-pass filter as only components with frequency $|\omega| \le \Omega$ survive the filtering process.

Reconstruction of the full signal f is generally not possible as information in components with frequency greater than Ω is unavailable. What is available is the signal

$$g = \mathcal{F}^{-1}\{\chi_{[-\Omega,\Omega]}\widehat{f}\}.$$

By the convolution theorem for Fourier transforms, one then has

$$g = \mathcal{F}^{-1}\{\chi_{[-\Omega,\Omega]}\} * f.$$

However,

$$\mathcal{F}^{-1}\{\chi_{[-\Omega,\Omega]}\}(t) = \frac{1}{2\pi}\int_{-\Omega}^{\Omega} e^{i\omega t}\,d\omega = \frac{\sin \Omega t}{\pi t}.$$

The reconstruction (or *extrapolation*) of the full signal f given the detected signal g requires the solution of the convolution equation

$$g(t) = \int_{-\infty}^{\infty} \frac{\sin\left(\Omega(t-\tau)\right)}{\pi(t-\tau)} f(\tau)\,d\tau.$$

The problem of extrapolating a band-limited signal is then seen to be mathematically the same as deblurring the effect of an instrument with the one-dimensional point spread function

$$k_\Omega(t) = \frac{\sin\left(\Omega t\right)}{\pi t}.$$

PET

CT scanning with X-rays is an instance of *transmission* tomography. A decade and a half prior to Cormack's publications on transmission tomography, an *emission* tomography technique, now known as PET (positron transmission tomography), was proposed [72]. In PET, a metabolically active tracer in the form of a positron-emitting isotope is injected into an area for which it has an affinity and taken up (metabolized) by an organ. The isotope emits positrons that immediately combine with free electrons in so-called annihilation events, which result in the ejection of two photons (γ-rays) along oppositely directed collinear rays. When a pair of detectors located on an array surrounding the body pick up the simultaneous arrival of two photons, one at each detector, respectively, an annihilation event is assumed to have taken place on the segment connecting the two detectors. In PET, the data collected from a very large number of such events is used to construct a two-dimensional tomographic slice of the isotope distribution. Because the uptake of the isotope is metabolically driven, PET is an effective tool for studying metabolism

giving it a diagnostic advantage over X-ray CT scanning. A combination of an X-ray CT scan with a PET scan provides the diagnostician anatomical information (distribution of attenuation coefficient) and physiological information (density of metabolized tracer isotope), respectively.

If $f : \mathbf{R^2} \to \mathbf{R}$ (we consider only a simplified version of 2D-PET) is the density of the metabolized tracer isotope, then the number of annihilations occurring along the coincidence line L connecting two detectors is proportional to the line integral

$$\int_L f \, ds.$$

That is, the observed counts of annihilation events are measured by the Radon transform of the density f. However, this does not take account of attenuation effects and can under represent features of deep-seated tissue. If the attenuation distribution is $\mu(\cdot, \cdot)$, and the pair of photons resulting from an annihilation event on the coincidence line L traverse oppositely directed rays L_+ and L_- of L, emanating from the annihilation site, then the detected attenuated signal takes the form

$$\begin{aligned} g &= \int_L e^{-\int_{L_+} \mu du} e^{-\int_{L_-} \mu du} f ds \\ &= e^{-\int_L \mu du} \int_L f ds. \end{aligned}$$

The model operator may now be viewed as a bivariate operator $K(\mu, f) = g$, in which the operator $K(\cdot, f)$ is nonlinear and the operator $K(\mu, \cdot)$ is linear. In soft tissue the attenuation coefficient is essentially zero, and therefore the solution of the inverse problem is accomplished by a Radon inversion of $K(0, \cdot)$. PET scans may be performed in combination with X-ray CT scans; the CT scan provides the attenuation coefficient, which may then be used in the model above to find the isotope density. See [52] for an extensive survey of emission tomography.

Some Mathematics for Inverse Problems

> ► Philosophy is written in that great book which ever lies before our gaze – I mean the universe .... The book is written in the mathematical language ... without which one wanders in vain through a dark labyrinth.
> Galileo Galilei

Hilbert space is a familiar environment that is rich enough for a discussion of the chief mathematical issues that are important in the theory of inverse problems. For the most part we restrict our attention to real Hilbert spaces. The inner product and associated norm will be symbolized by $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$, respectively:

$$\|x\| = \sqrt{\langle x, x\rangle}.$$

We assume the reader is familiar with the basic properties of inner product spaces (see, e.g., [18, Chap. I]), including the Cauchy–Schwarz inequality

$$|\langle x, y\rangle| \le \|x\|\|y\|.$$

A Hilbert space H is *complete*, that is, Cauchy sequences in H converge:

$$\text{if} \quad \lim_{n,m\to\infty} \|x_n - x_m\| = 0, \quad \text{then} \quad \|x_n - x\| \to 0,$$

for some $x \in H$. The smallest (in the sense of inclusion) Hilbert space that contains a given inner product space is known as the *completion* of the inner product space. (Every inner product space has a unique completion.)

The space, denoted $L^2[a, b]$, of measurable functions on an interval $[a, b]$ whose squares are Lebesgue integrable, with inner product

$$\langle f, g\rangle = \int_a^b f(t)g(t)\,dt,$$

is the prototypical example of a Hilbert space. The *Sobolev space* of order m, $H^{(m)}[a, b]$, is the completion with respect to the norm

$$\|f\|_0 = \left(\sum_{k=0}^{m} \|f^{(k)}\|_0^2\right)^{1/2},$$

associated with the inner product

$$\langle f, g\rangle_m = \sum_{k=0}^{m} \langle f^{(k)}, g^{(k)}\rangle_0,$$

of the space of functions having m continuous derivatives on $[a, b]$. Here $\langle\cdot,\cdot\rangle_0$ and $\|\cdot\|_0$ are the $L^2[a, b]$ norm and inner product; of course, $H^{(0)}[a, b] = L^2[a, b]$.

Two vectors x and y in a Hilbert space H are called *orthogonal*, denoted $x \perp y$, if $\langle x, y\rangle = 0$. The Pythagorean Theorem,

$$x \perp y \iff \|x + y\|^2 = \|x\|^2 + \|y\|^2,$$

is a key notion that suggests the transfer of familiar geometrical ideas from Euclidean space to Hilbert space. The *orthogonal complement* of a set S is the closed subspace

$$S^{\perp} = \{x \in H : x \perp s, \quad \text{for all} \quad s \in S\}.$$

It is not difficult to show that if S is a subspace, then $S^{\perp\perp} = \overline{S}$, where $\bar{S}$ is the *closure* of S, that is, the smallest closed subspace that contains S. A closed subspace S of a Hilbert space H engenders a Cartesian decomposition of H, symbolized by $H = S \oplus S^{\perp}$, meaning that each $x \in H$ has a unique representation of the form $x = x_1 + x_2$, where $x_1 \in S$ is the projection of x onto S:

$$\|x - x_1\| = \inf\{\|x - y\| : y \in S\},$$

and similarly x_2 is the projection of x onto $S^{\perp}$. The projection of a vector x onto a closed subspace S is denoted by $P_S x$.

A set of mutually orthogonal vectors each of which has unit norm is called an *orthonormal* set. An orthonormal set S is *complete* if $S^{\perp} = \{0\}$. A *complete orthonormal system* for a Hilbert space is a sequence of vectors in H, which is complete and orthonormal. For example, $\{\sin n\pi t : n = 1, 2, 3, \ldots\}$ is a complete orthonormal system for the Hilbert space $L^2[0, 1]$. Each vector $x \in H$ has a convergent Fourier expansion in terms of a complete orthonormal system $\{\varphi_n\}_{n=1}^{\infty}$ for H:

$$x = \sum_{n=1}^{\infty} \langle x, \varphi_n \rangle \varphi_n,$$

of which *Parseval's identity* is an immediate consequence

$$\|x\|^2 = \sum_{n=1}^{\infty} |\langle x, \varphi_n \rangle|^2.$$

Weak Convergence

"Weak" notions are crucial to our development of mathematics for inverse problems. Suppose the class of functions of interest forms a real Hilbert space H with an inner product $\langle \cdot, \cdot \rangle$ and associated norm $\| \cdot \|$. A *functional* is a mapping from H to $\mathbf{R}$. It is helpful to think of a functional as a measurement on elements of H. Proportionality and additivity are natural features of most measuring processes. A functional $F : H \to \mathbf{R}$ with these features, that is, satisfying

$$F(\alpha x + \beta y) = \alpha F(x) + \beta F(y)$$

where α and β are scalars and $x, y \in H$, is called a *linear functional*. Another common and highly desirable feature of a measurement process is *continuity*: elements of H which are nearly the same should result in measurements that are nearly the same. In mathematical terms, a functional F is continuous if, as $n \to \infty$,

$$\|x_n - x\| \to 0 \quad \text{implies} \quad |F(x_n) - F(x)| \to 0.$$

For example, if the Hilbert space is $L^2[0, T]$, the space of square integrable functions on $[0, T]$, then the average value functional,

$$F(x) = \frac{1}{T}\int_0^T x(\tau)\, d\tau,$$

is a continuous linear functional. (This is an immediate consequence of the Cauchy–Schwarz inequality.)

The Riesz Representation Theorem characterizes continuous linear functionals on a Hilbert space:

A continuous linear functional F on H has a unique representation of the form

$$F(x) = \langle x, \varphi\rangle$$

for some $\varphi \in H$.

This result is so fundamental that it is worthwhile to sketch a micro-proof. We may assume that F is not identically zero (otherwise, take $\varphi = 0$), and hence there is a $z \in H$, with $F(z) = 1$, which is orthogonal to the closed subspace

$$N = \{x \in H : F(x) = 0\}.$$

Then $x - F(x)z \in N$ for all $x \in H$, and hence $\varphi = z/\|z\|^2$ fits the bill

$$0 = \langle x - F(x)z, z/\|z\|^2\rangle = \langle x, \varphi\rangle - F(x).$$

Any two distinct vectors in H are distinguishable by some measurement in the form of a continuous linear functional. Indeed, if $\langle x - y, \varphi\rangle = 0$ for all $\varphi \in H$, then $x = y$ (set $\varphi = x - y$). However, it is possible for a sequence of vectors $\{x_n\}$, which does not converge in H to any vector, nevertheless to be ultimately indistinguishable from some vector x by bounded linear functionals. This is the idea of weak convergence. We say that $\{x_n\}$ converges *weakly* to x, symbolized $x_n \rightharpoonup x$, if $\langle x_n, \varphi\rangle \to \langle x, \varphi\rangle$ for every $\varphi \in H$. The simple identity

$$\|x - x_n\|^2 = \langle x - x_n, x - x_n\rangle = \|x\|^2 + \|x_n\|^2 - 2\langle x_n, x\rangle$$

shows that if $x_n \rightharpoonup x$ and $\|x_n\| \to \|x\|$, then x_n *converges strongly* to x, that is, $\|x_n - x\| \to 0$.

In a Hilbert space, every sequence of vectors whose norms are uniformly bounded has a subsequence that is weakly convergent (e.g., [18], p. 205). We note that any complete orthonormal system $\{\varphi_n\}$ converges weakly to zero, for by Parseval's identity

$$\sum_n |\langle x, \varphi_n\rangle|^2 = \|x\|^2,$$

and hence $\langle x, \varphi_n\rangle \to 0$ as $n \to \infty$ for each $x \in H$.

A set is called weakly closed if it contains the weak limit of every weakly convergent sequence of vectors in the set. Hilbert spaces have the following feature (see, e.g., [30]), which is fundamental in the theory of optimization:

Suppose C is a weakly closed convex subset of a Hilbert space H. For each $x \in H$, there is a unique vector $P_C(x) \in C$ such that

$$\|x - P_C(x)\| = \inf\{\|x - u\| : u \in C\}.$$

$P_C(x)$ is called the metric projection of x onto C. It can be shown that a closed convex set is also weakly closed.

Linear Operators

A bounded linear operator from a Hilbert space H_1 into a Hilbert space H_2 is a mapping $K : H_1 \to H_2$, which is linear, $K(\alpha x + \beta y) = \alpha Kx + \beta Ky$, and for which the number

$$\|K\| = \sup\{\|Kx\|/\|x\| : x \neq 0\}$$

is finite. Note that we have used the same symbol for the norm in each of the spaces; this generally will be our practice in the sequel. If K is a bounded linear operator, then K is (uniformly) continuous since

$$\|Kx - Ky\| = \|K(x - y)\| \leq \|K\|\|x - y\|.$$

For our purposes, the most cogent example of a bounded linear operator is an integral operator $K : L^2[a, b] \to L^2[c, d]$ of the form

$$Kf(t) = \int_a^b k(t, s) f(s)\, ds, \quad c \leq t \leq d, \tag{7}$$

where $k(\cdot, \cdot) \in L^2([c, d] \times [a, b])$ is called the *kernel* of the integral operator. The kernel is called *degenerate* if it has the form

$$k(t, s) = \sum_{j=1}^{m} T_j(t) S_j(s)$$

where the T_j and the S_j are each linearly independent sets of functions of a single variable. In this case the range, $R(K)$, of the operator K is the finite-dimensional subspace

$$R(K) = \text{span}\{T_j : j = 1, \ldots, m\}$$

and

$$Kf(t) = \sum_{j=1}^{m} \langle k(t, \cdot), f \rangle T_j(t),$$

where $\langle \cdot, \cdot \rangle$ is the $L^2[a, b]$ inner product.

The *adjoint* of a bounded linear operator $K : H_1 \to H_2$ is the bounded linear operator $K^* : H_2 \to H_1$, which satisfies

$$\langle Kx, y\rangle = \langle x, K^*y\rangle$$

for all $x \in H_1$ and $y \in H_2$. For example, by changing the order of integration, one can see that the adjoint of the integral operator (7) is

$$K^*g(s) = \int_c^d k(u,s)g(u)\,du.$$

A bounded linear operator $K : H \to H$ is called *self-adjoint* if $K^* = K$. The *null-space* of a bounded linear operator $K : H_1 \to H_2$ is the closed subspace

$$N(K) = \{x \in H_1 : Kx = 0\}.$$

Note that $N(K^*K) = N(K)$ for if $x \in N(K^*K)$, then

$$0 = \langle K^*Kx, x\rangle = \langle Kx, Kx\rangle = \|Kx\|^2.$$

There are fundamental relationships between the null-space and the range

$$R(K) = \{Kx : x \in H_1\},$$

of a linear operator and its adjoint. In fact, $y \in R(K)^\perp$ if and only if

$$0 = \langle Kx, y\rangle = \langle x, K^*y\rangle$$

for all $x \in H_1$, and hence $R(K)^\perp = N(K^*)$. It follows that $\overline{R(K)} = R(K)^{\perp\perp} = N(K^*)^\perp$. Replacing K by K^* in these relations (noting that $K^{**} = K$), we obtain the four fundamental relationships:

$$R(K)^\perp = N(K^*),\ \overline{R(K)} = N(K^*)^\perp$$

$$R(K^*)^\perp = N(K),\ \overline{R(K^*)} = N(K)^\perp.$$

Examples in a previous section have highlighted the unstable nature of solutions of inverse problems. This instability is conveniently phrased in terms of linear operators that are *unbounded*. Unbounded linear operators are typically defined only on restricted subspaces of the Hilbert space. For example, $L^2[0,\pi]$ contains discontinuous, and hence nondifferentiable, functions. But the differentiation operator may be defined on the proper subspace of $L^2[0,\pi]$ consisting of differentiable functions with derivatives in $L^2[0,\pi]$. This differentiation operator is unbounded since

$$\left\| \frac{1}{n} \sin n^2 t \right\|^2 = \frac{\pi}{2n^2} \to 0 \quad \text{as} \quad n \to \infty,$$

while

$$\left\| \frac{d}{dt} \left(\frac{1}{n} \sin n^2 t \right) \right\|^2 = \frac{\pi}{2} n^2 \to \infty \quad \text{as} \quad n \to \infty.$$

The reverse diffusion process is also governed by an unbounded operator. As seen in (5), the operator L, which maps the temperature distribution $g(x) = u(x, T)$ to the initial temperature distribution $f(x) = u(x, 0)$, is defined on the subspace

$$\mathcal{D}(L) = \left\{ g \in L^2[0, \pi] : \sum_{n=1}^{\infty} e^{2kn^2 T} |b_n|^2 < \infty \right\},$$

where

$$b_n = \frac{2}{\pi} \int_0^{\pi} g(s) \sin ns \, ds.$$

L is unbounded because the functions $\varphi_m(s) = \sin ms$ reside in $\mathcal{D}(L)$ and satisfy $\|\varphi_m\|^2 = \pi/2$, but by the orthogonality relationships,

$$L\varphi_m = e^{km^2 T} \varphi_m,$$

and hence $\|L\varphi_m\|^2 = e^{2km^2 T} \pi/2 \to \infty$ as $m \to \infty$.

Compact Operators and the SVD

A bounded linear operator $K : H_1 \to H_2$ of the form

$$Kx = \sum_{j=1}^{r} \langle x, v_j \rangle u_j, \tag{8}$$

where $\{u_j\}_{j=1}^r$ is a linearly independent set of vectors in H_2 and $\{v_j\}_{j=1}^r$ is a set of vectors in H_1, is called an operator of finite rank (with rank $= r$). For example, an integral operator on $L^2[a, b]$ with a degenerate kernel is an operator of finite rank. Finite rank operators transform weakly convergent sequences into strongly convergent sequences: if $x_n \rightharpoonup x$, then

$$Kx_n = \sum_{j=1}^{r} \langle x_n, v_j \rangle u_j \to \sum_{j=1}^{r} \langle x, v_j \rangle u_j = Kx.$$

More generally, a linear operator is called *compact* if it enjoys this weak-to-strong continuity, that is, if $x_n \rightharpoonup x$ implies that $Kx_n \to Kx$. In terms of our metaphor of bounded linear functionals as measurements, one could say that if the linear operator K modeling an inverse problem is compact, and if all measurements on a sequence of functions $\{x_n\}$ are ultimately indistinguishable from the corresponding measurements on x, then the model values Kx_n are ultimately indistinguishable from Kx. That is, causes, which are ultimately indistinguishable by linear measurement processes, result in effects that are ultimately indistinguishable. It is therefore not surprising that compact operators occur frequently in models of linear inverse problems.

Erhard Schmidt's theory of singular functions, now called the singular value decomposition (SVD), is the most versatile and effective tool for the analysis of compact linear operators. (The SVD has been rediscovered several times in various contexts; for the curious history of the SVD, see [65].) The SVD extends the representation (8) in a particularly useful way. A singular system $\{v_j, u_j; \mu_j\}_{j=1}^{\infty}$ for a compact linear operator K bundles a complete orthonormal system $\{v_j\}_{j=1}^{\infty}$ for $N(K)^{\perp}$, consisting of eigenvectors of K^*K; a complete orthonormal system $\{u_j\}_{j=1}^{\infty}$ for $N(K^*)^{\perp} = \overline{R(K)}$, consisting of eigenvectors of KK^*; and a sequence of positive numbers μ_j, called singular values of K. The singular values and singular vectors are tied together by the relationships

$$Kv_j = \mu_j u_j, \quad K^*u_j = \mu_j v_j, \quad j = 1, 2, 3, \ldots.$$

Every compact linear operator has a singular system, and the action of the operator may be expressed in terms of the SVD as

$$Kx = \sum_{j=1}^{\infty} \mu_j \langle x, v_j \rangle u_j \tag{9}$$

In case K has finite rank r, this sum terminates at $j = r$, and otherwise

$$\mu_j \to 0, \quad \text{as} \quad j \to \infty.$$

We shall see that this fact is singularly important in the analysis of inverse problems.

As an example of the SVD of a nonself-adjoint compact operator, consider the integral operator $K : L^2[0, \pi] \to L^2[0, \pi]$ defined by

$$(Kf)(t) = \int_0^{\pi} h(t, u) f(u)\, du$$

where

$$h(t, u) = \begin{cases} 1, & 0 \le u \le t \\ 0, & t < u \le \pi. \end{cases}$$

One can verify that a singular system $\{v_j, u_j; \mu_j\}_{j=1}^{\infty}$ for this operator is

$$v_j(t) = \sqrt{\frac{2}{\pi}} \cos\left(\frac{2j+1}{2}t\right), \quad u_j(s) = \sqrt{\frac{2}{\pi}} \sin\left(\frac{2j+1}{2}s\right), \quad \mu_j = \frac{2}{2j+1}.$$

A compact operator has closed range if and only if it has finite rank. This follows from the SVD and the open mapping theorem (e.g., [18], p. 166). Indeed, if K is compact and $R(K)$ is closed, then the restricted operator $K : N(K)^{\perp} \to R(K)$ is one-to-one and onto, and hence has a bounded inverse. That is, there is a positive number m such that $\|Kx\| \geq m\|x\|$ for all $x \in N(K)^{\perp}$. But then, by (9),

$$\mu_j = \mu_j \|u_j\| = \|Kv_j\| \geq m\|v_j\| = m > 0,$$

and hence K has only finitely many singular values for otherwise $\mu_j \to 0$. This result is highly significant in inverse theory for it says that finite rank linear models, when pushed too far toward the limiting case of an operator of infinite rank, will inevitably result in instability.

In 1910, Emil Picard [60] established a criterion that characterizes the existence of solutions of an equation of the first kind

$$Kx = y, \tag{10}$$

where K is a compact linear operator. Picard's criterion plays a role in inverse theory analogous to that which the Fredholm alternative plays for integral equations of the second kind.

Since $\{v_j\}$ is a complete orthonormal system for $N(K)^{\perp}$, the series

$$\sum_{j=1}^{\infty} |\langle x, v_j \rangle|^2$$

is convergent (and equals $\|P_{N(K)^{\perp}} x\|^2$). However, if $y = Kx \in R(K)$, then

$$\langle x, v_j \rangle = \mu_j^{-1} \langle x, K^* u_j \rangle = \mu_j^{-1} \langle Kx, u_j \rangle = \mu_j^{-1} \langle y, u_j \rangle,$$

and so

$$\sum_{j=1}^{\infty} \mu_j^{-2} |\langle y, u_j \rangle|^2 < \infty$$

is a necessary condition for $y \in R(K)$.

On the other hand, this condition guarantees that the series

$$x = \sum_{j=1}^{\infty} \mu_j^{-1} \langle y, u_j \rangle v_j \tag{11}$$

is convergent in $R(K)^{\perp}$ and the singular value relations show that $Kx = P_{N(K^*)^{\perp}} y$. Taken together these results establish the *Picard Criterion*:

$$y \in R(K) \Leftrightarrow y \in N(K^*)^{\perp} \quad \text{and} \quad \sum_{j=1}^{\infty} \mu_j^{-2} |\langle y, u_j \rangle|^2 < \infty. \tag{12}$$

If y satisfies Picard's criterion, then $y = Kx$ where x is given by (11).

The Moore–Penrose Inverse

If $y \notin R(K)$, then the equation $Kx = y$ has no solution, but this should not prevent one from doing the best one can to try to solve the problem. Perhaps the best that can be done is to seek a vector u that is as near as possible to serving as a solution. A vector $u \in H_1$ that minimizes the quadratic functional

$$F(x) = \|Kx - y\|^2$$

is called a *least squares* solution of $Kx = y$. It is not hard to see that a least squares solution exists if and only if y belongs to the dense subspace $R(T) + R(T)^{\perp}$ of H_2. Also, as the geometry suggests, u is a least squares solution if and only if $y - Ku \in R(K)^{\perp} = N(K^*)$, and hence least squares solutions are vector v that satisfy the so-called normal equation

$$K^* K v = K^* y. \tag{13}$$

Furthermore, the solution set of (13) is closed and convex and therefore contains a unique vector nearest to the origin (i.e., of smallest norm), say $v^{\dagger}$. This smallest norm least squares solution $v^{\dagger}$ lies in $N(K)^{\perp}$, for otherwise $Pv^{\dagger} \neq 0$, where P is the orthogonal projector onto $N(K)$. The Pythagorean theorem then gives

$$\|v^{\dagger}\|^2 = \|v^{\dagger} - Pv^{\dagger}\|^2 + \|Pv^{\dagger}\|^2.$$

But, since $K^* K P v^{\dagger} = 0$, this implies that $v^{\dagger} - Pv^{\dagger}$ is a least squares solution with norm smaller than that of $v^{\dagger}$. This contradiction ensures that $v^{\dagger} \in N(K)^{\perp}$.

The operator $K^{\dagger} : \mathcal{D}(K^{\dagger}) \to N(K)^{\perp}$, which associates with each y in the dense subspace $\mathcal{D}(K^{\dagger}) = R(K) + R(K)^{\perp}$ of H_2 the unique minimum norm least squares solution $K^{\dagger} y \in N(K)^{\perp}$ of the equation $Kx = y$, is called the Moore–Penrose generalized inverse of K. (E. H. Moore died when Roger (now Sir Roger) Penrose was an infant; [5] tells the story of how the names of both men came to be associated with the generalized inverse.)

If K is a compact linear operator with SVD $\{v_j, u_j; \mu_j\}$ and $y \in \mathcal{D}(K^\dagger)$, then the vector

$$\sum_{j=1}^{\infty} \frac{1}{\mu_j} \langle y, u_j \rangle v_j$$

is well defined by Picard's criterion, resides in $N(K)^\perp$, and is a least squares solution. Therefore,

$$K^\dagger y = \sum_{j=1}^{\infty} \frac{1}{\mu_j} \langle y, u_j \rangle v_j. \tag{14}$$

The operator $K^\dagger$ so defined is linear, but it is unbounded (unless K has finite rank) since

$$\|u_n\| = 1, \quad \text{but} \quad \|K^\dagger u_n\| = 1/\mu_n \to \infty \quad \text{as} \quad n \to \infty. \tag{15}$$

There is an immense literature on generalized inverses; see [54] for a start.

Alternating Projection Theorem

Draw a pair of intersecting lines. Take a point at random in the plane spanned by the lines, and project it onto the first line. Then project that point onto the second line, and continue in this manner projecting alternately onto each line in turn. It soon becomes apparent that the sequence of points so generated zigzags and converges to the point that is common to both lines. In 1933, von Neumann showed the same behavior for two closed subspaces S_1 and S_2 of a Hilbert space H. Namely, for each $x \in H$,

$$(P_{S_2} P_{S_1})^n x \to P_{S_2 \cap S_1} x \quad \text{as} \quad n \to \infty,$$

where P_W stands for the orthogonal projector onto the closed subspace W. This result extends easily to the case where S_1 and S_2 are translates of closed subspaces, that is, closed affine sets (see, e.g., [18, 35] for proofs). In fact, a modification of the method, due to Boyle and Dykstra (see [18], p. 213), provides a sequence that converges to the metric projection onto the intersection of a finite collection of closed convex sets.

Stefan Kaczmarz [48] developed an alternating projection algorithm, independently of von Neumann, for approximating solutions of underdetermined systems of linear algebraic equations. (See [57] concerning Kaczmarz's early and tragic demise.) A solution $\vec{x} \in \mathbf{R}^n$ of a system of m linear equations in n unknowns with coefficient matrix A and right-hand side $\vec{b}$ lies in the intersection of the hyperplanes

$$\pi_i = \{\vec{x} \in \mathbf{R}^n : \langle \vec{a}_i, \vec{x} \rangle = b_i\}, \quad i = 1, 2, \ldots, m,$$

where $\vec{a}_i$ is the ith row vector of A. Kaczmarz's algorithm, which consists of successively and cyclically projecting onto these hyperplanes, produces a sequence of vectors that converges to that vector in the intersection of the hyperplanes, which is nearest to the initial approximation (see [20] for a complete treatment of the method). Kaczmarz's method has come to be known as ART (the algebraic reconstruction technique) in the tomography community.

4 Numerical Methods

> ► All of exact science is dominated by the idea of approximation.
> Bertrand Russell

Tikhonov Regularization

The unboundedness of the operator $K^\dagger$, displayed in (15), is a fundamental challenge when solving linear inverse problems of the form $Kx = y$. This unboundedness is manifested as instability when the data vector y contains errors, which is always the case in practical circumstances as the data result from observation and measurement. Small errors in high-order singular components $\langle y, u_n \rangle$ (n large) will be magnified by the factor $1/\mu_n$ in the representation (14), resulting in large deviations in the computed solution. Such instabilities in numerical solutions were noticed from the very beginning of the use of digital computers to solve linear inverse problems (see [31] for examples and references). The development of theoretical strategies to mitigate this instability is known as *regularization theory*.

One way to stabilize the solution process is to restrict the notion of solution. Tikhonov's classic result [66] of 1943 is an instance of this idea. In that paper Tikhonov treated the inverse problem of determining the spatial distribution of a uniform star-shaped mass lying below the horizontal surface from measurements of the gravitational potential on the surface. He showed that the inverse problem becomes well posed if the forward operator is restricted to a certain compact set. Another approach is to modify the forward operator itself without a restriction on its domain. In what has come to be known as Tikhonov regularization, the notion of solution is generalized to the minimum norm least squares solution, which is unstable, but a stable approximation to this generalized solution, depending on a *regularization parameter*, is constructed.

The idea of Tikhonov regularization may be introduced from either an algebraic or a variational viewpoint. Algebraically, the method, in its simplest form, consists in replacing the normal equation (13) with the second kind equation

$$K^*Kv + \alpha v = K^*y, \tag{16}$$

where α is a positive parameter. The key point is that the problem of solving (16) is *well posed*. Indeed,

$$\|x\|^2\|K^*K+\alpha I\| \geq \langle (K^*K+\alpha I)x, x\rangle = \|Kx\|^2+\alpha\|x\|^2 \geq \alpha\|x\|^2,$$

and hence $(K^*K+\alpha I)^{-1}$ is a bounded linear operator; in fact, $\|(K^*K+\alpha I)^{-1}\| \leq 1/\alpha$. The significance of this fact is that for fixed $\alpha > 0$, the approximation

$$x_\alpha = (K^*K+\alpha I)^{-1}K^*y \tag{17}$$

depends continuously on y. Specifically, suppose y^δ is an observed version of y satisfying $\|y-y^\delta\| \leq \delta$, and let x_α^δ be the approximation formed using this approximate data, that is,

$$x_\alpha^\delta = (K^*K+\alpha I)^{-1}K^*y^\delta.$$

From the SVD we have

$$x_\alpha - x_\alpha^\delta = \sum_{j=1}^{\infty} \frac{\mu_j}{\mu_j^2+\alpha}\langle y-y^\delta, u_j\rangle v_j,$$

and hence

$$\begin{aligned}\|x_\alpha - x_\alpha^\delta\|^2 &= \sum_{j=1}^{\infty} \frac{\mu_j^2}{\mu_j^2+\alpha}\frac{1}{\mu_j^2+\alpha}|\langle y-y^\delta, u_j\rangle|^2 \\ &\leq \tfrac{1}{\alpha}\textstyle\sum_{j=1}^{\infty}|\langle y-y^\delta, u_j\rangle|^2 \leq \delta^2/\alpha.\end{aligned} \tag{18}$$

If the minimum norm least squares solution $K^\dagger y$ satisfies the *source condition* $K^\dagger y = K^*Kw$, for some $w \in H_1$, then one can show that

$$K^\dagger y - x_\alpha = \alpha(K^*K+\alpha I)^{-1}K^*Kw$$

and hence

$$\|K^\dagger y - x_\alpha\|^2 = \alpha^2\sum_{j=1}^{\infty}\left(\frac{\mu_j^2}{\mu_j^2+\alpha}\right)^2|\langle w, v_j\rangle|^2 \leq \alpha^2\|w\|^2. \tag{19}$$

Combining this with (18), we see that if $K^\dagger y \in R(K^*K)$, then

$$\|x_\alpha^\delta - K^\dagger y\| \leq \delta/\sqrt{\alpha} + O(\alpha).$$

Therefore, an a priori choice of the regularization parameter of the form

$$\alpha = \alpha(\delta) = C\delta^{2/3}, \tag{20}$$

yields a convergence rate of the form

$$\left\| x_{\alpha(\delta)}^{\delta} - K^{\dagger} y \right\| = O(\delta^{2/3}). \tag{21}$$

This two thirds power rate is an asymptotic "brick wall" for Tikhonov regularization in the sense that it is impossible to uniformly improve it to a $o(\delta^{2/3})$ rate unless the compact operator K has finite rank (see [27]). Roughly speaking, this says that the best one can hope for is $2m$-digit accuracy in the solution if there is $3m$-digit accuracy in the data.

The Tikhonov approximation (17) has a variational characterization that is useful in both theoretical analysis and computational implementation. The equation (16) that characterizes the Tikhonov approximation is the Euler equation for the functional

$$F_{\alpha}(\cdot; y) = \|K \cdot - y\|^2 + \alpha \|\cdot\|^2, \tag{22}$$

and hence the approximation (17) is a global minimizer of (22). This opens the possibility of applying standard optimization techniques for calculating the Tikhonov approximation. Next we illustrate the usefulness of the variational characterization in a convergence analysis for an a posteriori selection technique for the regularization parameter known as Morozov's discrepancy principle.

The a priori parameter selection criterion (20) is of theoretical interest as it gives information on the order of magnitude of the regularization parameter that can be expected to result in a convergent procedure. However, a posteriori methods of choosing the regularization parameter that depend on the actual progress of the computations would be expected to lead to more satisfactory results. Morozov's *discrepancy principle* [53] is the earliest parameter choice strategy of this type. Morozov's idea (which was presaged by Phillips [31, 59]) is to choose the regularization parameter in such a way that the size of the residual $\left\| K x_{\alpha}^{\delta} - g^{\delta} \right\|$ is equal to error level in the data:

$$\left\| K x_{\alpha}^{\delta} - y^{\delta} \right\| = \delta. \tag{23}$$

It should be recognized that this condition contains some "slack" as δ, the bound for the data error, might not be tight. Nevertheless, this choice is not only possible, but it leads to a convergent procedure, as we now show.

If $\|y^{\delta}\| > \delta$, that is, there is more signal than noise in the data, and if $y \in R(K)$, then there is a unique positive parameter α satisfying (23). To see this, we use the SVD

$$\left\| Kx_\alpha^\delta - y^\delta \right\|^2 = \sum_{j=1}^{\infty} \left(\frac{\alpha}{\mu_j^2 + \alpha} \right)^2 |\langle y^\delta, u_j \rangle|^2 + \|Py^\delta\|^2 \tag{24}$$

where P is the orthogonal projector of H_2 onto $R(K)^\perp$. From this we see that the real function

$$\psi(\alpha) = \left\| Kx_\alpha^\delta - y^\delta \right\|$$

is a continuous, strictly increasing function of α satisfying (since $Py = 0$)

$$\lim_{\alpha \to 0^+} \psi(\alpha) = \|Py^\delta\| = \|Pg^\delta - Pg\| \leq \|y^\delta - y\| \leq \delta$$

and

$$\lim_{\alpha \to \infty} \psi(\alpha) = \|y^\delta\| > \delta.$$

The intermediate value theorem then guarantees a unique $\alpha = \alpha(\delta)$ satisfying (23).

We now show that the choice $\alpha(\delta)$ as given by the discrepancy method (23) leads to a regular scheme for approximating $K^\dagger y$:

$$x_{\alpha(\delta)}^\delta \to K^\dagger y \quad \text{as} \quad \delta \to 0.$$

To this end it is sufficient to show that for any sequence $\delta_n \to 0$, there is a subsequence, which we will denote by $\{\delta_k\}$, that satisfies $x_{\alpha(\delta_k)}^{\delta_k} \to K^\dagger y$. The argument relies on the following previously discussed facts: norm-bounded sequences contain a weakly convergent subsequence, and weak convergence along with convergence of the norms implies strong convergence.

We assume that $y \in R(K)$ and that $K : H_1 \to H_2$ is a compact linear operator, and we let $x = K^\dagger y$. That is, x is the unique vector in $N(K)^\perp$ satisfying $Kx = y$.

The variational characterization of the Tikhonov approximation $x_{\alpha(\delta)}^\delta$ as the global minimizer of the quadratic functional $F_\alpha(\cdot\,; y^\delta)$ (see (22)) implies that

$$F_{\alpha(\delta)}\left(x_{\alpha(\delta)}^\delta; y^\delta \right) \leq F_{\alpha(\delta)}(x; y^\delta),$$

that is,

$$\begin{aligned} \delta^2 + \alpha(\delta) \left\| x_{\alpha(\delta)}^\delta \right\|^2 &= \left\| Kx_{\alpha(\delta)}^\delta - y^\delta \right\|^2 + \alpha(\delta) \left\| x_{\alpha(\delta)}^\delta \right\|^2 \\ &\leq F_{\alpha(\delta)}(x) = \|y - y^\delta\|^2 + \alpha(\delta)\|x\|^2 \\ &\leq \delta^2 + \alpha(\delta)\|x\|^2 \end{aligned}$$

and hence $\left\|x^{\delta}_{\alpha(\delta)}\right\| \leq \|x\|$. Therefore, for any sequence $\delta_n \to 0$, there is a subsequence $\delta_k \to 0$ with $x^{\delta_k}_{\alpha(\delta_k)} \rightharpoonup w$ for some w. But

$$x^{\delta}_{\alpha(\delta)} = K^*(KK^* + \alpha(\delta)I)^{-1}y^{\delta} \in R(K^*) \subseteq N(K)^{\perp}$$

and $N(K)^{\perp}$ is weakly closed, and so $w \in N(K)^{\perp}$. Furthermore,

$$\left\|Kx^{\delta_k}_{\alpha(\delta_k)} - y^{\delta_k}\right\| \to 0$$

and hence $Kx^{\delta_k}_{\alpha(\delta_k)} \to y$. But as K is compact, $Kx^{\delta_k}_{\alpha(\delta_k)} \to Kw$, and it follows that $Kw = y$ and $w \in N(K)^{\perp}$, that is, $w = x$. Since $\left\|x^{\delta_k}_{\alpha(\delta_k)}\right\| \leq \|x\|$, we then have

$$\|x\|^2 = \lim_{k\to\infty}\left\langle x^{\delta_k}_{\alpha(\delta_k)}, x\right\rangle \leq \underline{\lim}_{k\to\infty}\left\|x^{\delta_k}_{\alpha(\delta_k)}\right\| \cdot \|x\|$$

and therefore

$$\|x\| \leq \underline{\lim}_{k\to\infty}\left\|x^{\delta_k}_{\alpha(\delta_k)}\right\| \leq \overline{\lim}_{k\to\infty}\left\|x^{\delta_k}_{\alpha(\delta_k)}\right\| \leq \|x\|.$$

So we have shown that $x^{\delta_k}_{\alpha(\delta_k)} \rightharpoonup x$ and $\left\|x^{\delta_k}_{\alpha(\delta_k)}\right\| \to \|x\|$, and hence $x^{\delta_k}_{\alpha(\delta_k)} \to x$, completing the proof.

It can be shown that, under the source condition $x \in R(K^*)$, Tikhonov's method with parameter choice by the discrepancy principle (23) achieves an asymptotic order of accuracy $O(\sqrt{\delta})$; however, a swifter rate of $o(\sqrt{\delta})$ is generally impossible except in the case when K has finite rank [26]. Engl and Gfrerer (see [19, Chap. 4]) have developed a modification of the discrepancy principle that achieves the optimal order of convergence.

Our sketch of the basic theory of Tikhonov regularization has assumed that the regularization functional, which augments the least square objective functional $\|K\cdot - y\|^2$, is (the square of) a norm. (Note however that while the same symbol is used for the norm in each of the spaces H_1 and H_2, these norms may be distinct. In his original paper [67], Tikhonov used a Sobolev norm on the solution space and an L^2 norm on the data space.) Phillips [59], in a paper that barely predates that of Tikhonov, used a regularizing semi-norm – the L^2 norm of the second derivative. In all of these cases, the equation characterizing the regularized approximation is linear. However, certain non-quadratic regularizing functionals, leading to nonlinear problems for determining the regularized approximation, are found to be effective in imaging science. Of particular note is the total variation, or TV, functional:

$$\mathrm{TV}(u) = \int_{\Omega} |\nabla u|,$$

where $u : \Omega \subseteq \mathbf{R}^n \to \mathbf{R}$. Regularization now consists of minimizing the augmented least squares functional

$$F_\alpha(u) = \|Ku - y\|^2_{L^2(\Omega)} + \alpha \mathrm{TV}(u),$$

where K is the identity operator for denoising problems, while for deblurring problems, K is the blurring operator associated with a known point spread function. A full exposition may be found in [62].

Iterative Regularization

Ordinary Tikhonov regularization consists in minimizing the functional

$$F_\alpha(z) = \|Kz - y\|^2 + \alpha \|z\|^2$$

for a range of positive regularization parameters α. In iterated Tikhonov regularization, $\alpha > 0$ is *fixed*, an initial approximation x_0 is selected (we take $x_0 = 0$ for simplicity; a general initial approximation requires only small modifications to the arguments), and successive approximations are updated by a multistage optimization scheme in which the nth approximation is chosen to minimize the functional

$$F_n(z) = \|Kz - y\|^2 + \alpha \|z - x_{n-1}\|^2, \quad n = 1, 2, 3, \ldots \tag{25}$$

This results in the iterative method

$$x_n = (K^*K + \alpha I)^{-1}(\alpha x_{n-1} + K^* y), \quad n = 1, 2, 3, \ldots.$$

The conventional proof of the convergence of iterated Tikhonov regularization uses spectral theory. (See [41] for the more general case of nonstationary iterated Tikhonov regularization.) However, the convergence of the method is also an immediate consequence of the alternating projection theorem, as we now show.

Let $\mathcal{H}$ be the product Hilbert space $H_1 \times H_2$ with norm $|\cdot|$ given by

$$|(x, y)|^2 = \|y\|^2 + \alpha \|x\|^2,$$

where α is a fixed positive constant. Note that the graph

$$\mathcal{G} = \{(x, Kx) : x \in H_1\}$$

is a closed subspace of $\mathcal{H}$. For a given $y \in H_2$, let

$$L_y = \{u \in H_1 : Ku = Py\},$$

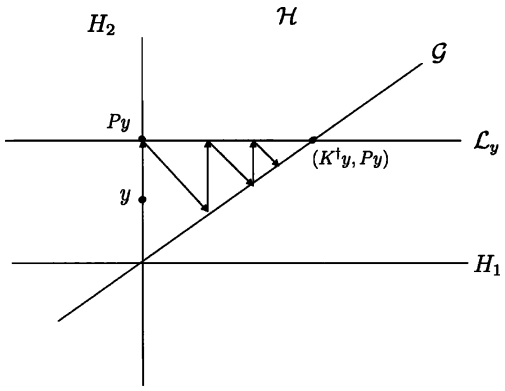


Fig. 3 The geometry of iterated regularization

where P is the orthogonal projector of H_2 onto $\overline{R(K)}$, be the set of least squares solutions of $Kx = y$. One sees that $y \in \mathcal{D}(K^\dagger)$ if and only if L_y is nonempty. If $\mathcal{L}_y = H_1 \times \{Py\}$, then $\mathcal{L}_y$ is a closed affine set in the Hilbert space $\mathcal{H}$, and $y \in \mathcal{D}(K^\dagger) \Leftrightarrow \mathcal{L}_y \cap \mathcal{G} \neq \emptyset$. Furthermore,

$$\mathcal{P}_{\mathcal{L}_y \cap \mathcal{G}}(0, y) = \left(K^\dagger y, Py\right),$$

where $\mathcal{P}_W$ stands for the metric projector of $\mathcal{H}$ onto a closed convex set $W \subseteq \mathcal{H}$.

From the variational characterization (25), one sees that

$$\mathcal{P}_{\mathcal{L}_y}(x_0, Kx_0) = \mathcal{P}_{\mathcal{L}_y}(0, 0) = (0, Py)$$

and $\mathcal{P}_{\mathcal{G}}(0, Py) = (x_1, Kx_1)$; therefore $(x_1, Kx_1) = \mathcal{P}_{\mathcal{G}}\mathcal{P}_{\mathcal{L}_y}(x_0, Kx_0)$, and generally

$$(x_n, Kx_n) = \mathcal{P}_{\mathcal{G}}\mathcal{P}_{\mathcal{L}_y}(x_{n-1}, Kx_{n-1}) = \ldots = (\mathcal{P}_{\mathcal{G}}\mathcal{P}_{\mathcal{L}_y})^n(0, y).$$

This process of projecting in alternate fashion in the space $\mathcal{H}$ is illustrated in Fig. 3.

The alternating projection theorem then gives

$$(x_n, Kx_n) \to \mathcal{P}_{\mathcal{L}_y \cap \mathcal{G}}(0, y) = (K^\dagger y, Py), \quad \text{as} \quad n \to \infty.$$

If x_n^δ are defined as in (25), with y replaced by y^δ, then it is not difficult to see that

$$\left\|x_n - x_n^\delta\right\| \leq \sqrt{n}\|y - y^\delta\|,$$

and hence if $\|y - y^\delta\| \leq \delta$ and $n = n(\delta) \to \infty$ as $\delta \to 0$, in such as manner that $\sqrt{n(\delta)}\delta \to 0$, then $x_{n(\delta)}^\delta \to K^\dagger x$.

There are many other iterative methods for regularization of ill-posed problems (see [19, 49]). Perhaps the simplest is based on the observation that for any $\lambda > 0$, the subspace $N(K)^{\perp}$ is invariant under the mapping

$$F(z) = (I - \lambda K^* K)x + \lambda K^* y,$$

and $K^{\dagger}y$ is the unique fixed point of F in $N(K)^{\perp}$. Furthermore, if $0 < \lambda < 1/\|K^*K\|$, then F is a contraction and hence the iterative method

$$x_{n+1} = (I - \lambda K^* K)x_n + \lambda K^* y \tag{26}$$

converges to $K^{\dagger}y$ for any $x_0 \in N(K)^{\perp}$. This method was studied for Fredholm integral equations of the first kind by Landweber and independently by Fridman. It has since become known as Landweber iteration [51]. For this method one can show easily that if $\|y - y^{\delta}\| \leq \delta$, and x_n^{δ} represents the approximation obtained by (26) with y replaced by y^{δ}, then $x_{n(\delta)}^{\delta} \to K^{\dagger}y$, if $\sqrt{n(\delta)}\delta \to 0$.

The theory of Landweber iteration has been developed for nonlinear operators, including a stopping criterion based on the discrepancy principle, by Hanke et al. [42] (see also [49]). See [40] for a very useful survey of iterative regularization methods.

Discretization

> ► . . . numerical precision is the very soul of science . . .
> D'Arcy Wentworth Thompson

The preceding discussion of regularization methods took place in the context of general (infinite-dimensional) Hilbert space. However, practical numerical computations are necessarily finitary. Passing from general elements in an infinite-dimensional space to finitely represented approximations involves a process of *discretization*. Discretization of an ill-posed problem can lead to a well-posed finite-dimensional problem; however, this discretized version generally will be *ill conditioned*. Regularization methods are meant to address this problem. There are two approaches to constructing computable regularizations of linear inverse problems; one could handle the ill posedness by first regularizing the infinite-dimensional problem and then discretizing the result, or one could discretize the original problem and then regularize the resulting ill-conditioned finite-dimensional problem. We give a couple of examples of discretized regularizations of the former type. (For results on discretized versions of general regularization methods, see [34].)

A key point in the theoretical convergence analysis of regularization methods is the interplay between the regularization parameter and the error properties of the data. For example, assuming a source condition of the form $K^{\dagger}y \in R(K^*K)$, the

balancing of the rate $O(\alpha)$ for the infinite-dimensional Tikhonov approximation x_α using "clean" data with the stability bound $\delta/\sqrt{\alpha}$ for the approximation using noisy data leads to the optimal rate $O(\delta^{2/3})$ found in (21). To obtain an overall convergence rate with respect to the error in data for discretized approximations, it is necessary to juggle three balls: a theoretical convergence rate, a measure of the quality of the discretization, and a stability bound. In both of the cases we consider, the measure of quality of the discretization will be denoted by γ_m. In our first example, γ_m measures how well a given finite-dimensional subspace $V_m \subseteq H_1$ supports the operator K, specifically,

$$\gamma_m = \|K(I - P_m)\|,$$

where P_m is the orthogonal projector of H_1 onto V_m. The smaller γ_m is, the better the subspace V_m supports the operator K. In the second example, the discretization of a regularized version of a Fredholm integral equation of the first kind is accomplished by applying a quadrature method to the iterated kernel that generates the operator K^*K. In this case, γ_m measures the quality of this quadrature. In both examples it is shown that it is theoretically possible to match the optimal rate $O(\delta^{2/3})$ established for the infinite-dimensional approximation in (21).

The variational characterization (22) immediately suggests a Ritz approach to discretization, namely, minimization of the Tikhonov functional over a finite-dimensional subspace. Note that the global minimum x_α of the functional $F_\alpha(\cdot\,; y)$ on H_1 may be characterized by the condition

$$\langle Kx_\alpha - Kx, Kv\rangle + \alpha\langle x_\alpha, v\rangle = 0, \quad \text{for all } v \in H_1, \tag{27}$$

where $x = K^\dagger y$. The bilinear form defined on H_1 by

$$q(u, v) = \langle Ku, Kv\rangle + \alpha\langle u, v\rangle$$

is an inner product on H_1, and (27) may be succinctly expressed in terms of this inner product as

$$q(x_\alpha - x, v) = 0 \quad \text{for all } v \in H_1.$$

Suppose that $\{V_m\}_{m=1}^\infty$ is a sequence of finite-dimensional subspaces of H_1 satisfying

$$V_1 \subseteq V_2 \subseteq V_3 \subseteq \ldots \subseteq H_1 \quad \text{and} \quad \overline{\cup_{m=1}^\infty V_m} = H_1.$$

The minimizer $x_{\alpha,m}$ of $F_\alpha(\cdot\,; y)$ over the finite-dimensional subspace V_m satisfies

$$q(x_{\alpha,m} - x, v_m) = 0 \quad \text{for all } v_m \in V_m,$$

and hence

$$q(x_\alpha - x_{\alpha,m}, v_m) = 0 \quad \text{for all } v_m \in V_m.$$

In other words, $x_{\alpha,m} = \mathcal{P}_m x_\alpha$, where $\mathcal{P}_m$ is the projector of H_1 onto V_m, which is orthogonal in the sense of the inner product $q(\cdot\,,\cdot)$.

If $|\cdot|_q$ denotes the norm on H_1 associated with the inner product $q(\cdot\,,\cdot)$, that is,

$$|z|_q^2 = \|Kz\|^2 + \alpha\|z\|^2,$$

then, by the characteristic property of projectors,

$$|x_\alpha - x_{\alpha,m}|_q^2 = |x_\alpha - \mathcal{P}_m x_\alpha|_q^2 \le |x_\alpha - P_m x_\alpha|_q^2,$$

where P_m is the projector of H_1 onto V_m associated with the (original) inner product on H_1. But then (since projectors are idempotent),

$$\begin{aligned}\alpha\|x_\alpha - x_{\alpha,m}\|^2 \le |x_\alpha - x_{\alpha,m}|_q^2 &\le \|Kx_\alpha - KP_m x_\alpha\|^2 + \alpha\|(I - P_m)x_\alpha\|^2 \\ &= \|K(I - P_m)^2 x_\alpha\|^2 + \alpha\|(I - P_m)x_\alpha\|^2 \\ &\le (\gamma_m + \alpha)\|(I - P_m)x_\alpha\|^2,\end{aligned}$$

where

$$\gamma_m = \|K(I - P_m)\|.$$

Therefore,

$$\|x_\alpha - x_{\alpha,m}\| \le \sqrt{1 + \gamma_m/\alpha} \quad \|(I - P_m)x_\alpha\|.$$

If $K^\dagger y$ satisfies the source condition $x = K^\dagger y \in R(K^*K)$, say, $x = K^*Kw$, then

$$(I - P_m)x_\alpha = (I - P_m)K^*(KK^* + \alpha I)^{-1}KK^*Kw,$$

and hence

$$\|(I - P_m)x_\alpha\| \le \gamma_m\|Kw\|.$$

If $\gamma_m = O(\alpha_m)$, then we find from (19),

$$\|K^\dagger y - x_{\alpha,m}\| = O(\alpha_m).$$

In the case of approximate data y^δ satisfying $\|y - y^\delta\| \le \delta$, one can show, using arguments of the same type as above, that a stability bound of the same form as (18) holds for the finite-dimensional approximations:

$$\left\|x_{\alpha,m} - x_{\alpha,m}^\delta\right\| \le \delta/\sqrt{\alpha}.$$

Taking these results together, we see that if $K^\dagger y \in R(K^*K)$ and $\alpha_m = \alpha_m(\delta)$ is chosen in such a way that $\alpha_m = C\delta^{2/3}$ and $\gamma_m = O(\alpha_m)$, then the finite-dimensional approximations achieve the optimal order of convergence:

$$\left\|K^\dagger y - x_{\alpha(m),m}^\delta\right\| = O(\delta^{2/3}).$$

Quadrature is another common discretization technique. If a linear inverse problem is expressed as a Fredholm integral equation of the first kind

$$y(s) = \int_a^b k(s,t)x(t)dt, \quad c \le s \le d,$$

mapping functions $x \in L^2[a,b]$ to function $y \in L^2[c,d]$, then the Tikhonov approximation x_α is the solution of the well-posed Fredholm integral equation of the second kind

$$\int_c^d k(u,s)y(u)du = \alpha x_\alpha(s) + \int_a^b \tilde{k}(s,t)x_\alpha(t)dt, \quad a \le s \le b,$$

where the iterated kernel $\tilde{k}(\cdot,\cdot)$ is given by

$$\tilde{k}(s,t) = \int_c^d k(u,s)k(u,t)du, \quad a \le s,t \le b.$$

If a convergent quadrature scheme with positive weights $\left\{w_j^{(m)}\right\}_{j=1}^m$, and nodes $\left\{u_j^{(m)}\right\}_{j=1}^m$, is applied to the iterated kernel, a degenerate kernel

$$\tilde{k}_m(s,t) = \sum_{j=1}^m w_j^{(m)} k\left(u_j^{(m)}, s\right) k\left(u_j^{(m)}, t\right)$$

results, converting the infinite-dimensional Tikhonov problem into the finite rank problem

$$\alpha x_{\alpha,m} + \tilde{K}_m x_{\alpha,m} = K^* y \tag{28}$$

where

$$\tilde{K}_m z = \sum_{j=1}^{m} w_j^{(m)} \langle k_j, z \rangle k_j \quad \text{and} \quad k_j(s) = k\left(u_j^{(m)}, s\right).$$

The problem (28) is equivalent to an $m \times m$ linear algebraic system with a unique solution. The convergence of the approximations resulting from this finite system to the infinite-dimensional Tikhonov approximation $x_\alpha \in L^2[a,b]$ depends on the number

$$\gamma_m = \|\tilde{K}_m - K^* K\|.$$

If $\alpha = \alpha(m) \to 0$ as $m \to \infty$ and $\gamma_m = O(\alpha(m))$, then it can be shown that $x_{\alpha(m),m} \to K^\dagger y$. Furthermore, a stability bound of the form $O(\delta/\sqrt{\alpha})$ holds under appropriate conditions, and one can show that the optimal rate $O(\delta^{2/3})$ is achievable if the parameters governing the finite-dimensional approximations are appropriately related [29]. A much more extensive analysis along these lines is carried out in [11]. For more on numerical methods for discrete inverse problems, see [43, 70].

5 Conclusion

> ► Eventually, we reach the dim boundary ...
> There, we measure shadows ...
> Edwin Hubble

The first book devoted exclusively to the mathematical theory of inverse and ill-posed problems was that of Tikhonov and Arsenin [68]. Kirsch [50] is a fine treatment of the general theory of inverse problems, and Engl et al. [19] is the best comprehensive presentation of the theory of regularization for inverse and ill-posed problems. Other useful books on the general topic are [46] and [69]. A number of books and survey articles treat inverse theory in a specific context. Some of the areas treated include astronomy [16], engineering [45], geophysics [58], imaging [6,7,9,10,20,44,55,62], mathematical physics [24], oceanography [4,73], parameter estimation [3], indirect measurement [2], and vibration analysis [23].

Cross-References

- ▸ Expansion Methods
- ▸ Inverse Scattering
- ▸ Iterative Solution Methods
- ▸ Large-Scale Inverse Problems in Imaging
- ▸ Numerical Methods and Applications in Total Variation Image Restoration
- ▸ Regularization Methods for Ill-Posed Problems
- ▸ Tomography
- ▸ Total Variation in Imaging
- ▸ Variational Methods in Shape Analysis

References

1. Ambarzumian, V.: On the derivation of the frequency function of space velocities of the stars from the observed radial velocities. Mon. Not. R. Astron. Soc. Lond. **96**, 172–179 (1936)
2. Anderssen, R.S.: Inverse problems: a pragmatist's approach to the recovery of information from indirect measurements. Aust. N.Z. Ind. Appl. Math. J. **46**, 588–622 (2004)
3. Aster, R., Borchers, B., Thurber, C.: Parameter Estimation and Inverse Problems. Elsevier, Boston (2005)
4. Bennett, A.: Inverse Modeling of the Ocean and Atmosphere. Cambridge University Press, Cambridge (2002)
5. Ben-Israel, A.: The Moore of the Moore penrose inverse. Electron. J. Linear Algebr. **9**, 150–157 (2002)
6. Bertero, M., Boccacci, P.: Introduction to Inverse Problems in Imaging. IOP, London (1998)
7. Bonilla, L. (ed.): Inverse Problems and Imaging. LNM 1943. Springer, Berlin (2008)
8. Carasso, A., Sanderson, J., Hyman, J.: Digital removal of random media image degradations by solving the diffusion equation backwards in time. SIAM J. Numer. Anal. **15**, 344–367 (1978)
9. Chalmond, B.: Modeling and Inverse Problems in Image Analysis. Springer, New York (2003)
10. Chan, T.F., Shen, J.: Image Processing and Analysis. SIAM, Philadelphia (2005)
11. Chen, Z., Xu, Y., Yang, H.: Fast collocation methods for solving ill-posed integral equations of the first kind. Inverse Probl. **24**, 065007(21) (2008)
12. Cormack, A.: Representation of a function by its line integrals, with some radiological applications I. J. Appl. Phys. **34**, 2722–2727 (1963)
13. Cormack, A.: Representation of a function by its line integrals, with some radiological applications II. J. Appl. Phys. **35**, 2908–2912 (1964)
14. Cormack, A.: Computed tomography: some history and recent developments. In: Shepp, L.A. (ed.) Computed Tomography. Proceedings of Symposia in Applied Mathematics, vol. 27, pp. 35–42. American Mathematical Society, Providence (1983)
15. Courant, R., Hilbert, D.: Methods of Mathematical Physics. Partial Differential Equations, vol. 2. Interscience, New York (1962)
16. Craig, I., Brown, J.: Inverse Problems in Astronomy. Adam Hilger, Bristol (1986)
17. Deans, S.R.: The Radon Transform and Some of Its Applications. Wiley, New York (1983)
18. Deutsch, F.: Best Approximation in Inner Product Spaces. Springer, New York (2001)
19. Engl, H.W., Hanke, M., Neubauer, A.: Regularization of Inverse Problems. Kluwer, Dordrecht (1996)
20. Epstein, C.L.: Introduction to the Mathematics of Medical Imaging. Pearson Education, Upper Saddle River (2003)
21. Galilei, G.: Sidereus Nuncius (Trans.: van Helden, A.). University of Chicago Press, Chicago, 1989 (1610)

22. Gates, E.: Einstein's Telescope. W.W. Norton, New York (2009)
23. Gladwell, G.M.L.: Inverse Problems in Vibration. Martinus Nijhoff, Dordrecht (1986)
24. Glasko, V.: Inverse Problems of Mathematical Physics (Trans.: Bincer, A. (Russian)). American Institute of Physics, New York (1984)
25. Goldberg, R.R.: Fourier Transforms. Cambridge University Press, Cambridge (1961)
26. Groetsch, C.W.: Comments on Morozov's Discrepancy Principle. In: Hämmerlin, G., Hoffmann, K.-H. (eds.) Improperly Posed Problems and Their Numerical Treatment, pp. 97–104. Birkhäuser, Basel (1983)
27. Groetsch, C.W.: On the asymptotic order of convergence of Tikhonov regularization. J. Optim. Theory Appl. **41**, 293–298 (1983)
28. Groetsch, C.W.: The Theory of Tikhonov Regularization for Fredholm Equations of the First Kind. Pitman, Boston (1984)
29. Groetsch, C.W.: Convergence analysis of a regularized degenerate kernel method for Fredholm integral equations of the first kind. Integr. Equ. Oper. Theory **13**, 67–75 (1990)
30. Groetsch, C.W.: Inverse Problems in the Mathematical Sciences. Vieweg, Braunschweig (1993)
31. Groetsch, C.W.: The delayed emergence of regularization theory. Bollettino di Storia delle Scienze Matematiche **23**, 105–120 (2003)
32. Groetsch, C.W.: Nascent function concepts in Nova Scientia. Int. J. Math. Educ. Sci. Technol. **35**, 867–875 (2004)
33. Groetsch, C.W.: Extending Halley's problem. Math. Sci. **34**, 4–10 (2009)
34. Groetsch, C.W., Neubauer, A.: Regularization of ill-posed problems: optimal parameter choice in finite dimensions. J. Approx. Theory **58**, 184–200 (1989)
35. Groetsch, C.W.: Stable Approximate Evaluation of Unbounded Operators. LNM 1894. Springer, New York (2007)
36. Grosser, M.: The Discovery of Neptune. Harvard University Press, Cambridge (1962)
37. Hadamard, J.: Sur les problèmes aux dériveès partielles et leur signification physique. Princet. Univ. Bull. **13**, 49–52 (1902)
38. Hadamard, J.: Lectures on Cauchy's Problems in Linear Partial Differential Equations. Yale University Press, New Haven (1923). (Reprinted by Dover, New York, 1952)
39. Halley, E.: A discourse concerning gravity, and its properties, wherein the descent of heavy bodies, and the motion of projects is briey, but fully handled: together with the solution of a problem of great use in gunnery. Philos. Trans. R. Soc. Lond. **16**, 3–21 (1686)
40. Hanke, M.: Iterative regularization techniques in image reconstruction. In: Colton, D. et al. (eds.) Surveys on Solution Methods for Inverse Problems, pp. 35–52. Springer, Vienna (2000)
41. Hanke, M., Groetsch, C.W.: Nonstationary iterated Tikhonov regularization. J. Optim. Theory Appl. **98**, 37–53 (1998)
42. Hanke, M., Neubauer, A., Scherzer, O.: A convergence analysis of Landweber iteration for nonlinear ill-posed problems. Numer. Math. **72**, 21–37 (1995)
43. Hansen, P.C.: Rank Deficient and Discrete Ill-Posed Problems. SIAM, Philadelphia (1997)
44. Hansen, P.C., Nagy, J., O'Leary, D.: Deblurring images: matrices, spectra, and filtering. SIAM, Philadelphia (2006)
45. Hensel, E.: Inverse Theory and Applications for Engineers. Prentice-Hall, Englewood Cliffs (1991)
46. Hofmann, B.: Regularization for Applied Inverse and Ill-Posed Problems. Teubner, Leipzig (1986)
47. Joachimstahl, F.: Über ein attractionsproblem. J. für die reine und angewandte Mathematik **58**, 135–137 (1861)
48. Kaczmarz, S.: Angenäherte Auflösung von Systemen linearer Gleichungen. Bulletin International de l'Academie Polonaise des Sciences, Cl. d. Sc. Mathém. A, pp. 355–357 (1937)
49. Kaltenbacher, B., Neubauer, A., Scherzer, O.: Iterative Regularization Methods for Nonlinear Ill-Posed Problems. Walter de Gruyter, Berlin (2008)
50. Kirsch, A.: An Introduction to the Mathematical Theory of Inverse Problems. Springer, New York (1993)

51. Landweber, L.: An iteration formula for Fredholm integral equations of the first kind. Am. J. Math. **73**, 615–624 (1951)
52. Lewitt, R.M., Matej, S.: Overview of methods for image reconstruction from projections in emission computed tomography. Proc. IEEE **91**, 1588–1611 (2003)
53. Morozov, V.A.: On the solution of functional equations by the method of regularization. Sov. Math. Dokl. **7**, 414–417 (1966)
54. Nashed, M.Z. (ed.): Generalized Inverses and Applications. Academic, New York (1976)
55. Natterer, F., Wübberling, F.: Mathematical Methods in Image Reconstruction. SIAM, Philadelphia (2001)
56. Newbury, P., Spiteri, R.: Inverting gravitational lenses. SIAM Rev. **44**, 111–130 (2002)
57. Parks, P.C., Kaczmarz, S.: Int. J. Control **57**, 1263–1267 (1895–1939)
58. Parker, R.L.: Geophysical Inverse Theory. Princeton University Press, Princeton (1994)
59. Phillips, D.L.: A technique for the numerical solution of certain integral equations of the first kind. J. Assoc. Comput. Mach. **9**, 84–97 (1962)
60. Picard, E.: Sur un théorème général relatif aux équations intégrales de premiére espéce et sur quelques probl_emes de physique mathématique. Rendiconti del Cicolo Matematico di Palermo **29**, 79–97 (1910)
61. Radon, J.: Über die Bestimmung von Funktionen durch ihre Integralwerte längs gewisser Mannigfaltigkeiten. Berichte über die Verhandlungen der Königlich Sächsischen Gesellshaft der Wissenschaften zur Leipzig **69**, 262–277 (1917)
62. Scherzer, O., Grasmair, M., Grossauer, H., Haltmeier, M., Lenzen, F.: Variational Methods in Imaging. Springer, New York (2009)
63. Sheehan, W., Kollerstrom, N., Waff, C.: The case of the pilfered planet: did the British steal Neptune? Sci. Am. **291**(6), 92–99 (2004)
64. Shepp, L.A. (ed.): Computed Tomography. Proceedings of Symposia in Applied Mathematics, vol. 27. American Mathematical Society, Providence (1983)
65. Stewart, G.W.: On the early history of the singular value decomposition. SIAM Rev. **35**, 551–566 (1993)
66. Tikhonov, A.N.: On the stability of inverse problems. Dokl. Akad. Nau. SSSR **39**, 176–179 (1943)
67. Tihonov (Tikhonov), A.N.: Solution of incorrectly formulated problems and the regularization method. Sov. Math. Dokl. **4**, 1035–1038 (1963)
68. Tikhonov, A.N., Arsenin, V.Y.: Solutions of Ill-Posed Problems. Winston & Sons, Washington, DC (1977)
69. Uhlmann, G. (ed.): Inside Out: Inverse Problems and Applications. Cambridge University Press, New York (2003)
70. Vogel, C.R.: Computational Methods for Inverse Problems. SIAM, Philadelphia (2002)
71. Wing, G.M.: A Primer on Integral Equations of the First Kind: The Problem of Deconvolution and Unfolding. SIAM, Philadelphia (1992)
72. Wrenn, F.R., Good, M.L., Handler, P.: The use of positron-emitting radioisotopes for the localization of brain tumors. Science **113**, 525–527 (1951)
73. Wunsch, C.: The Ocean Circulation Inverse Problem. Cambridge University Press, Cambridge (1996)

Large-Scale Inverse Problems in Imaging

Julianne Chung, Sarah Knepper, and James G. Nagy

Contents

1 Introduction 48
2 Background 49
Model Problems 49
Imaging Applications 50
3 Mathematical Modeling and Analysis 56
Linear Problems 56
Separable Inverse Problems 67
Nonlinear Inverse Problems 71
4 Numerical Methods and Case Examples 73
Linear Example: Deconvolution 74
Separable Example: Multi-Frame Blind Deconvolution 78
Nonlinear Example: Tomosynthesis 81
5 Conclusion 85
Cross-References 86
References 87

Abstract

Large-scale inverse problems arise in a variety of significant applications in image processing, and efficient regularization methods are needed to compute meaningful solutions. This chapter surveys three common mathematical models including a linear model, a separable nonlinear model, and a general nonlinear model. Techniques for regularization and large-scale implementations are con-

J. Chung (✉)
Virginia Tech, Blacksburg, VA, USA
e-mail: jmchung@vt.edu

S. Knepper • J.G. Nagy
Emory University, Atlanta, GA, USA
e-mail: nagy@mathcs.emory.edu

© Springer Science+Business Media New York 2015

O. Scherzer (ed.), *Handbook of Mathematical Methods in Imaging*,
DOI 10.1007/978-1-4939-0790-8_2

sidered, with particular focus on algorithms and computations that can exploit structure in the problem. Examples from image deconvolution, multi-frame blind deconvolution, and tomosynthesis illustrate the potential of these algorithms. Much progress has been made in the field of large-scale inverse problems, but many challenges still remain for future research.

1 Introduction

Powerful imaging technologies, including very large telescopes, synthetic aperture radar, medical imaging scanners, and modern microscopes, typically combine a device that collects electromagnetic energy (e.g., photons) with a computer that assembles the collected data into images that can be viewed by practitioners, such as scientists and doctors. The "assembling" process typically involves solving an *inverse problem*; that is, the image is reconstructed from indirect measurements of the corresponding object. Many inverse problems are also *ill-posed*, meaning that small changes in the measured data can lead to large changes in the solution, and special tools or techniques are needed to deal with this instability. In fact, because real data will not be exact (it will contain at least some small amount of noise or other errors from the data collection device), it is not possible to find the exact solution. Instead, a physically realistic approximation is sought. This is done by formulating an appropriate *regularized* (i.e., stabilized) problem, from which a good approximate solution can be computed.

Inverse problems are ubiquitous in imaging applications, including deconvolution (or, more generally, deblurring) [1, 51], superresolution (or image fusion) [18, 27], image registration [70], image reconstruction [74, 75], seismic imaging [31], inverse scattering [15], and radar imaging [17]. These problems are referred to as *large-scale* because they typically require processing a large amount of data (the number of pixels or voxels in the discretized image) and systems with a large (e.g., 10^9 for a 3D image reconstruction problem) number of equations. Mathematicians began to rigorously study inverse problems in the 1960s, and this interest has continued to grow over the past few decades due to applications in fields such as biomedical, seismic, and radar imaging; see, for example, [12, 28, 47, 49, 99] and the references therein.

We remark that the discussion in this chapter does not address some very important issues that can arise in PDE-based inverse problems, such as adjoints and proper meshing. Inverse problems such as these arise in important applications, including PDE parameter identification, seismic imaging, and inverse scattering; we refer those interested in these topics and applications to the associated chapters in this handbook and the references therein.

This chapter discusses computational approaches to compute approximate solutions of large-scale inverse problems. Mathematical models and some applications are presented in Sect. 2. Three basic models are considered: a general nonlinear model, a linear model, and a mixed linear/nonlinear model. Several regularization approaches are described in Sect. 3. Numerical methods that can be used to compute

approximate solutions for the three basic models, along with illustrative examples from specific imaging applications, are described in Sect. 4. Concluding remarks, including a partial list of open questions, are provided in Sect. 5.

2 Background

A mathematical framework for inverse problems is presented in this chapter, including model problems and imaging applications. Although only a limited number of imaging applications are considered, the model problems, which range from linear to nonlinear, are fairly general and can be used to describe many other applications. For more complete treatments of inverse problems and regularization, see [12, 28, 47, 49, 50, 99].

Model Problems

An inverse problem involves the estimation of certain quantities using information obtained from indirect measurements. A general mathematical model to describe this process is given by

$$\mathbf{b}_{\text{exact}} = F(\mathbf{x}_{\text{exact}}), \tag{1}$$

where $\mathbf{x}_{\text{exact}}$ denotes the exact (or ideal) quantities that need to be estimated and $\mathbf{b}_{\text{exact}}$ is used to represent perfectly measured (error-free) data. The function F is defined by the data collection process and is assumed known. Typically, it is assumed that F is defined on Hilbert spaces and that it is continuous and weakly sequentially closed [29].

Unfortunately, in any real application, it is impossible to collect error-free data, so a more realistic model of the data collection process is given by

$$\mathbf{b} = F(\mathbf{x}_{\text{exact}}) + \boldsymbol{\eta}, \tag{2}$$

where $\boldsymbol{\eta}$ represents noise and other errors in the measured data. The precise form of F depends on the application; the following three general problems are considered in this chapter:

- For linear problems, $F(\mathbf{x}) = \mathbf{A}\mathbf{x}$, where $\mathbf{A}$ is a linear operator. In this case, the data collection process is modeled as

$$\mathbf{b} = \mathbf{A}\mathbf{x}_{\text{exact}} + \boldsymbol{\eta},$$

  and the inverse problem is: given $\mathbf{b}$ and $\mathbf{A}$, compute an approximation of $\mathbf{x}_{\text{exact}}$.
- In some cases, $\mathbf{x}$ can be separated into two distinct components, $\mathbf{x}^{(\ell)}$ and $\mathbf{x}^{(n\ell)}$, with $F(\mathbf{x}) = F\left(\mathbf{x}^{(\ell)}, \mathbf{x}^{(n\ell)}\right) = A\left(\mathbf{x}^{(n\ell)}\right)\mathbf{x}^{(\ell)}$, where $\mathbf{A}$ is a linear operator defined

by $\mathbf{x}^{(n\ell)}$. That is, the data $\mathbf{b}$ depends linearly on $\mathbf{x}^{(\ell)}$ and nonlinearly on $\mathbf{x}^{(n\ell)}$. In this case, the data collection process is modeled as

$$\mathbf{b} = \mathbf{A}\left(\mathbf{x}_{\text{exact}}^{(n\ell)}\right)\mathbf{x}_{\text{exact}}^{(\ell)} + \eta,$$

and the inverse problem is: given $\mathbf{b}$ and the parametric form of $\mathbf{A}$, compute approximations of $\mathbf{x}_{\text{exact}}^{(n\ell)}$ and $\mathbf{x}_{\text{exact}}^{(\ell)}$.

- If the problem is not linear or separable, as described above, then the general nonlinear model,

$$\mathbf{b} = F(\mathbf{x}_{\text{exact}}) + \boldsymbol{\eta},$$

will be considered. In this case, the inverse problem is: given $\mathbf{b}$ and F, compute an approximation of $\mathbf{x}_{\text{exact}}$.

In most of what follows, it is assumed that the problem has been discretized, so $\mathbf{x}$, $\mathbf{b}$, and $\boldsymbol{\eta}$ are vectors, and $\mathbf{A}$ is a matrix. Depending on the constraints assumed and the complexity of the model used, problems may range from linear to fully nonlinear. This is true of the applications described in the next subsection.

Imaging Applications

Three applications in image processing that lead to inverse problems are discussed in this subsection. For each application, the underlying mathematical model is described, and some background for the problem is presented. The formulation of each of these problems results in linear, separable, and nonlinear inverse problems, respectively.

Image Deblurring and Deconvolution

In many important applications, such as when ground-based telescopes are used to observe objects in space, the observed image is degraded by blurring and noise. Although the blurring can be partially avoided by using sophisticated and expensive imaging devices, computational post-processing techniques are also often needed to further improve the resolution of the image. This post-processing is known as *image deblurring*. To give a precise mathematical model of image deblurring, suppose $x(t), t \in \mathcal{R}^d$, is a scalar function describing the true d-dimensional (e.g., for a plane image containing pixels, $d = 2$) image. Then the observed, blurred, and noisy image is given by

$$b(s) = \int_{\Omega} k(s,t)x(t)dt + \eta(s), \tag{3}$$

where $s \in \mathcal{R}^d$ and $\eta(s)$ represents additive noise. The kernel $k(s,t)$ is a function that specifies how the points in the image are distorted and is therefore called the point spread function (PSF). The inverse problem of image deblurring is: given k and b, compute an approximation of x. If the kernel has the property that $k(s,t) = k(s-t)$, then the PSF is said to be spatially invariant; otherwise, it is said to be spatially variant. In the spatially invariant case, the blurring operation, $\int k(s-t)x(t)dt$, is convolution, and thus the corresponding inverse problem is called *deconvolution*.

In a realistic problem, images are collected only at discrete points (pixels or voxels) and are only available in a finite bounded region. Therefore, one must usually work directly either with a semi-discrete model

$$b(s_j) = \int_\Omega k(s_j,t)x(t)dt + \eta_j \quad j = 1,\ldots,N$$

where N is the number of pixels or voxels in the observed image or with the fully discrete model

$$\mathbf{b} = \mathbf{A}\mathbf{x}_{\text{exact}} + \boldsymbol{\eta},$$

where $\mathbf{x}_{\text{exact}}, \mathbf{b}$, and $\boldsymbol{\eta}$ are vectors obtained by discretizing functions x, b, and η and $\mathbf{A}$ is a matrix that arises when approximating the integration operation with, for example, a quadrature rule. Moreover, a precise kernel representation of the PSF may not be known but instead must be constructed experimentally from the imaging system by generating images of "point sources." What constitutes a point source depends on the application. For example, in atmospheric imaging, the point source can be a single bright star [53]. In microscopy, the point source is typically a fluorescent microsphere having a diameter that is about half the diffraction limit of the lens [24]. For general motion blurs, the PSF is described by the direction (e.g., angle) and speed at which objects are moving [56].

For spatially invariant blurs, one point source image and appropriate boundary conditions are enough to describe the matrix $\mathbf{A}$. This situation has been well studied; algorithms to compute approximations of $\mathbf{x}$ can be implemented efficiently with fast Fourier transforms (FFT) or other trigonometric transforms [1, 51]. More recently, an approach has been proposed where the data can be transformed to the Radon domain so that computations can be done efficiently with, for example, wavelet filtering techniques [26].

Spatially variant blurs also occur in a variety of important applications. For example, in positron emission tomography (PET), patient motion during the relatively long scan times causes reconstructed images to be corrupted by nonlinear, nonuniform spatially variant motion blur [33, 84]. Spatially variant blurs also occur when the object and image coordinates are tilted relative to each other, as well as in X-ray projection imaging [100], lens distortions [65], and wave aberrations [65]. Moreover, it is unlikely that the blur is truly spatially invariant in any realistic application, especially over large image planes.

Various techniques have been proposed to approximately model spatially variant blurs. For example, in the case of patient motion in PET brain imaging, a motion detection device is used to monitor the position of the patient's head during the scan time. This information can then be used to construct a large sparse matrix $\mathbf{A}$ that models the motion blur. Other, more general techniques include coordinate transformation [68], image partitioning [93], and PSF interpolation [72, 73].

Multi-Frame Blind Deconvolution

The image deblurring problem described in the previous subsection assumes that the blurring operator, or PSF, is known. However, in most cases, only an approximation of the operator, or an approximation of parameters that defines the operator, is known. For example, as previously mentioned, the PSF is often constructed experimentally from the imaging system by generating images of point sources. In many cases, such approximations are fairly good and are used to construct the matrix $\mathbf{A}$ in the linear model. However, there are situations where it is not possible to obtain good approximations of the blurring operator, and it is necessary to include this knowledge in the mathematical model. Specifically, consider the general image formation model

$$\mathbf{b} = \mathbf{A}\left(\mathbf{x}_{\text{exact}}^{(n\ell)}\right)\mathbf{x}_{\text{exact}}^{(\ell)} + \boldsymbol{\eta} \tag{4}$$

where $\mathbf{b}$ is a vector representing the observed, blurred, and noisy image and $\mathbf{x}_{\text{exact}}^{(\ell)}$ is a vector representing the unknown true image to be reconstructed. $\mathbf{A}\left(\mathbf{x}_{\text{exact}}^{(n\ell)}\right)$ is an ill-conditioned matrix defining the blurring operator. For example, in the case of spatially invariant blurs, $\mathbf{x}_{\text{exact}}^{(n\ell)}$ could simply be the pixel (image space) values of the PSF. Or $\mathbf{x}_{\text{exact}}^{(n\ell)}$ could be a small set of parameters that define the PSF, such as with a Zernike polynomial-based representation [67]. In general, the number of parameters defining $\mathbf{x}_{\text{exact}}^{(n\ell)}$ is significantly smaller than the number of pixels in the observed image. As in the previous subsection, $\boldsymbol{\eta}$ is a vector that represents unknown additive noise in the measured data. The term *blind deconvolution* is used for algorithms that attempt to jointly compute approximations of $\mathbf{x}_{\text{exact}}^{(n\ell)}$ and $\mathbf{x}_{\text{exact}}^{(\ell)}$ from the separable inverse problem given by Eq. (4).

Blind deconvolution problems are highly underdetermined, which present many challenges to optimization algorithms that can easily become trapped in local minima. This difficulty has been well-documented; see, for example, [64, 67]. To address challenges of nonuniqueness, it may be necessary to include additional constraints, such as nonnegativity and bounds on the computed approximations $\mathbf{x}^{(n\ell)}$ and $\mathbf{x}^{(\ell)}$.

Multi-frame blind deconvolution (MFBD) [64, 67] reduces some of the nonuniqueness problems by collecting multiple images of the same object, but with different blurring operators. Specifically, suppose a set of (e.g., m) observed images of the same object are modeled as

$$\mathbf{b}_i = \mathbf{A}\left(\mathbf{x}_i^{(n\ell)}\right)\mathbf{x}_{\text{exact}}^{(\ell)} + \boldsymbol{\eta}_i, \quad i = 1, 2, \ldots, m. \tag{5}$$

Then, a general separable inverse problem of the form given by Eq. (4) can be obtained by setting

$$\mathbf{b} = \begin{bmatrix} \mathbf{b}_1 \\ \vdots \\ \mathbf{b}_m \end{bmatrix}, \quad \mathbf{x}_{\text{exact}}^{(n\ell)} = \begin{bmatrix} \mathbf{x}_1^{(n\ell)} \\ \vdots \\ \mathbf{x}_m^{(n\ell)} \end{bmatrix}, \quad \boldsymbol{\eta} = \begin{bmatrix} \boldsymbol{\eta}_1 \\ \vdots \\ \boldsymbol{\eta}_m \end{bmatrix}.$$

Although multiple frames reduce, to some extent, the nonuniqueness problem, they do not completely eliminate it. In addition, compared to single-frame blind deconvolution, there is a significant increase in the computational complexity of processing the large, multiple data sets.

There are many approaches to solving the blind and multi-frame blind deconvolution problem; see, for example, [14]. In addition, many other imaging applications require solving separable inverse problems, including super resolution (which is an example of image data fusion) [18, 27, 57, 76], the reconstruction of 3D macromolecular structures from 2D electron microscopy images of cryogenically frozen samples (Cryo-EM) [22, 35, 55, 66, 82, 88], and seismic imaging applications [40].

Tomosynthesis

Modern conventional X-ray systems that use digital technology have many benefits to the classical film X-ray systems, including the ability to obtain high-quality images with lower-dosage X-rays. The term "conventional" is used to refer to a system that produces a 2D projection image of a 3D object, as opposed to computed tomography (CT), which produces 3D images. Because of the inexpensive cost, low X-ray dosage, and ease of use, digital X-ray systems are widely used in medicine, from emergency rooms, to mammography, to dentistry.

Tomosynthesis is a technique that can produce 3D image information of an object using conventional X-ray systems [25]. The basic idea underlying tomosynthesis is that multiple 2D image projections of the object are taken at varying incident angles, and each 2D image provides different information about the 3D object. See Fig. 1 for an illustration of a typical geometry for breast tomosynthesis imaging. The relationship between the multiple 2D image projections and the 3D object can be modeled as a nonlinear inverse problem. Reconstruction algorithms that solve this inverse problem should be able to reconstruct any number of slices of the 3D object. Sophisticated approaches used for 3D CT reconstruction cannot be applied here because projections are only taken from a limited angular range, leaving entire regions of the frequency space unsampled. Thus, alternative approaches need to be considered.

The mathematical model described in this section is specifically designed for breast imaging and assumes a polyenergetic (i.e., multiple energy) X-ray source. It is first necessary to determine what quantity will be reconstructed. Although most X-ray projection models are derived in terms of the values of the attenuation coefficients for the voxels, it is common in breast imaging to interpret the voxels as a composition of adipose tissue, glandular tissue, or a combination of both [42].

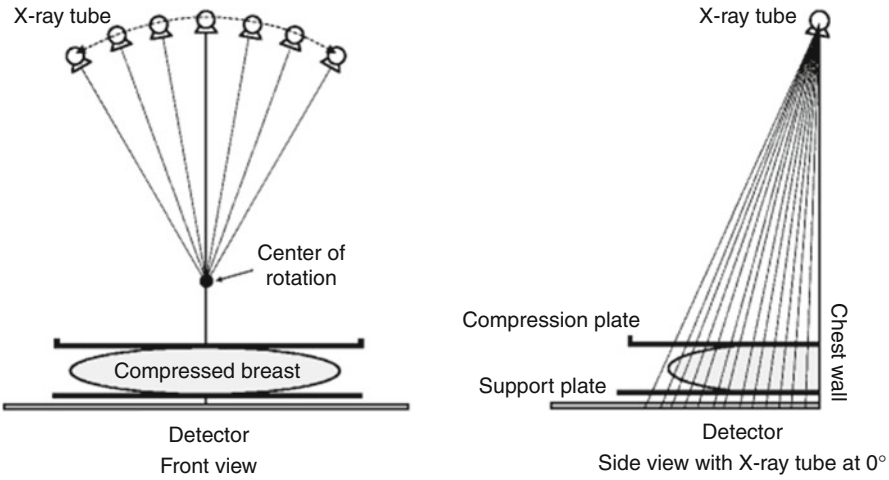


Fig. 1 Breast tomosynthesis example. Typical geometry of the imaging device used in breast imaging

Thus, each voxel of the object can be represented using the percentage glandular fraction, that is, the percentage of glandular tissue present in that voxel. If density or attenuation coefficient values are desired, then these can be obtained from the glandular fraction through a simple algebraic transformation.

Now assume that the 3D object is discretized into a regular grid of voxels and that each of the 2D projection images is discretized into a regular grid of pixels. Specifically, let N represent the number of voxels in the discretized 3D object and let M be the number of pixels in a discretized 2D projection image. In practice, N is on the order of a few billion and M is the order of a few million, depending on the size of the imaging detector. The energy-dependent linear attenuation coefficient for voxel $j = 1, 2, \ldots, N$ in the breast can be represented as

$$\mu(e)^{(j)} = s(e)x_{\text{exact}}^{(j)} + z(e),$$

where $x_{\text{exact}}^{(j)}$ represents the percentage glandular fraction in voxel j of the "true" object and $s(e)$ and $z(e)$ are known energy-dependent linear fit coefficients. This type of decomposition to reduce the number of degrees of freedom, which is described in more detail in [20], is similar to an approach used by De Man et al. [23] for CT, in which they express the energy-dependent linear attenuation coefficient in terms of its photoelectric component and Compton scatter component.

The projections are taken from various angles in a predetermined angular range, and the photon energies can be discretized into a fixed number of levels. Let there be n_θ angular projections and assume the incident X-ray has been discretized into n_e photon energy levels. In practice, a typical scan may have $n_\theta = 21$ and $n_e = 43$. For a particular projection angle, compute a monochromatic ray trace for one energy

level and then sum over all energies. Let $a^{(ij)}$ represent the length of the ray that passes through voxel j, contributing to pixel i. Then, the discrete monochromatic ray trace for pixel i can be represented by

$$\sum_{j=1}^{N} \mu(e)^{(j)} a^{(ij)} = s(e) \sum_{j=1}^{N} x_{\text{exact}}^{(j)} a^{(ij)} + z(e) \sum_{j=1}^{N} a^{(ij)}. \tag{6}$$

Using the standard mathematical model for transmission radiography, the ith pixel value for the θth noise-free projection image, incorporating all photon energies present in the incident X-ray spectrum, can be written as

$$b_{\theta}^{(i)} = \sum_{e=1}^{n_e} \varrho(e) \exp\left(- \sum_{j=1}^{N} \mu(e)^{(j)} a^{(ij)} \right), \tag{7}$$

where $\varrho(e)$ is a product of the current energy with the number of incident photons at that energy. To simplify notation, define A_{θ} to be an $M \times N$ matrix with entries $a^{(ij)}$. Then Eq. (6) gives the ith entry of the vector

$$s(e)\mathbf{A}_{\theta}\mathbf{x}_{\text{exact}} + z(e)\mathbf{A}_{\theta}\mathbf{1},$$

where $\mathbf{x}_{\text{exact}}$ is a vector whose jth entry is $x_{\text{exact}}^{(j)}$ and $\mathbf{1}$ is a vector of all ones. Furthermore, the θth noise-free projection image in vector form can be written as

$$\mathbf{b}_{\theta} = \sum_{e=1}^{n_e} \varrho(e) \exp(-[s(e)\mathbf{A}_{\theta}\mathbf{x}_{\text{exact}} + z(e)\mathbf{A}_{\theta}\mathbf{1}]), \tag{8}$$

where the exponential function is applied component-wise.

Tomosynthesis reconstruction is a nonlinear inverse problem where the goal is to approximate the volume, $\mathbf{x}_{\text{exact}}$, given the set of projection images from various angles, $\mathbf{b}_{\theta}, \theta = 1, 2, \ldots, n_{\theta}$. This can be put in the general nonlinear model

$$\mathbf{b} = F(\mathbf{x}_{\text{exact}}) + \boldsymbol{\eta},$$

where

$$\mathbf{b} = \begin{bmatrix} \mathbf{b}_1 \\ \vdots \\ \mathbf{b}_{\mathrm{n}_{\theta}} \end{bmatrix} \quad \text{and} \quad F(\mathbf{x}) = \begin{bmatrix} \sum_{e=1}^{n_e} \varrho(e) \exp(-[s(e)\mathbf{A}_1\mathbf{x} + z(e)\mathbf{A}_1\mathbf{1}]) \\ \vdots \\ \sum_{e=1}^{n_e} \varrho(e) \exp(-[s(e)\mathbf{A}_{\mathrm{n}_{\theta}}\mathbf{x} + z(e)\mathbf{A}_{\mathrm{n}_{\theta}}\mathbf{1}]) \end{bmatrix}$$

3 Mathematical Modeling and Analysis

A significant challenge when attempting to compute approximate solutions of inverse problems is that they are typically ill-posed. To be precise, in 1902 Hadamard defined a well-posed problem as one that satisfies the following requirements:

1. The solution is unique;
2. The solution exists for arbitrary data;
3. The solution depends continuously on the data.

Ill-posed problems, and hence most inverse problems, typically fail to satisfy at least one of these criteria. It is worth mentioning that this definition of an ill-posed problem applies to continuous mathematical models and not precisely to the discrete approximations used in computational methods. However, the properties of the continuous ill-posed problem are often carried over to the discrete problem in the form of a particular kind of ill-conditioning, making certain (usually high frequency) components of the solution very sensitive to errors in the measured data; this property is discussed in more detail for linear problems in section "Linear Problems." Of course, this may depend on the level of discretization; a coarsely discretized problem may not be very ill-conditioned, but it also may not bear much similarity to the underlying continuous problem.

Regularization is a term used to refer to various techniques that modify the inverse problem in an attempt to overcome the instability caused by ill-posedness. Regularization seeks to incorporate a priori knowledge into the solution process. Such knowledge may include information about the amount or type of noise, the smoothness or sparsity of the solution, or restrictions on the values the solution may obtain. Each regularization method also requires choosing one or more regularization parameters. A variety of approaches are discussed in this section.

The theory for regularizing linear problems is much more developed than it is for nonlinear problems. This is due, in large part, to the fact that the numerical treatment of nonlinear inverse problems is often highly dependent on the particular application. However, good intuition can be gained by first studying linear inverse problems.

Linear Problems

Consider the linear inverse problem

$$\mathbf{b} = \mathbf{A}\mathbf{x}_{\text{exact}} + \boldsymbol{\eta},$$

where $\mathbf{b}$ and $\mathbf{A}$ are known, and the aim is to compute an approximation of $\mathbf{x}_{\text{exact}}$. The linear problem is a good place to illustrate the challenges that arise when attempting

to solve large-scale inverse problems. In addition, some of the regularization methods and iterative algorithms discussed here can be used in, or generalized for, nonlinear inverse problems.

SVD Analysis

A useful tool in studying linear inverse problems is the singular value decomposition (SVD). Any $m \times n$ matrix $\mathbf{A}$ can be written as

$$\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T \tag{9}$$

where $\mathbf{U}$ is an $m \times m$ orthogonal matrix, $\mathbf{V}$ is an $n \times n$ orthogonal matrix, and $\boldsymbol{\Sigma}$ is an $m \times n$ diagonal matrix containing the singular values $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_{\min(m,n)} \geq 0$. If $\mathbf{A}$ is nonsingular, then an approximation of $\mathbf{x}_{\text{exact}}$ is given by the inverse solution

$$\mathbf{x}_{\text{inv}} = \mathbf{A}^{-1}\mathbf{b} = \sum_{i=1}^{n} \frac{\mathbf{u}_i^T \mathbf{b}}{\sigma_i}\mathbf{v}_i = \sum_{i=1}^{n} \underbrace{\frac{\mathbf{u}_i^T \mathbf{b}_{\text{exact}}}{\sigma_i}\mathbf{v}_i}_{\mathbf{x}_{\text{exact}}} + \sum_{i=1}^{n} \underbrace{\frac{\mathbf{u}_i^T \boldsymbol{\eta}}{\sigma_i}\mathbf{v}_i}_{\text{error}}$$

where $\mathbf{u}_i$ and $\mathbf{v}_i$ are the singular vectors of $\mathbf{A}$ (i.e., the columns of $\mathbf{U}$ and $\mathbf{V}$, respectively). As indicated above, the inverse solution is comprised of two components: $\mathbf{x}_{\text{exact}}$ and an error term. Before discussing algorithms to compute approximations of $\mathbf{x}_{\text{exact}}$, it is useful to study the error term.

For matrices arising from ill-posed inverse problems, the following properties hold:

- The matrix $\mathbf{A}$ is severely ill-conditioned, with the singular values σ_i decaying to zero without a significant gap to indicate numerical rank.
- The singular vectors corresponding to the small singular values tend to oscillate more (i.e., have higher frequency) than singular vectors corresponding to large singular values.
- The components $\left|\mathbf{u}_i^T \mathbf{b}_{\text{exact}}\right|$ decay on average faster than the singular values σ_i. This is referred to as the *discrete Picard condition* [49].

The first two properties imply that the high-frequency components of the error term are highly magnified by division of small singular values. The computed inverse solution is dominated by these high-frequency components and is in general a very poor approximation of $\mathbf{x}_{\text{exact}}$. However, the third property suggests that there is hope of reconstructing some information about $\mathbf{x}_{\text{exact}}$; that is, an approximate solution can be obtained by reconstructing components corresponding to the large singular values and filtering out components corresponding to small singular values.

Regularization by SVD Filtering

The SVD filtering approach to regularization is motivated by observations made in the previous subsection. That is, by filtering out components of the solution corresponding to the small singular values, a reasonable approximation of $\mathbf{x}_{\text{exact}}$ can be computed. Specifically, an SVD-filtered solution is given by

$$\mathbf{x}_{\text{filt}} = \sum_{i=1}^{n} \phi_i \frac{\mathbf{u}_i^T \mathbf{b}}{\sigma_i} \mathbf{v}_i, \tag{10}$$

where the *filter factors*, ϕ_i, satisfy $\phi_i \approx 1$ for large σ_i, and $\phi_i \approx 0$ for small σ_i.

That is, the large singular value components of the solution are reconstructed, while the components corresponding to the small singular values are filtered out. Different choices of filter factors lead to different methods. Some examples include:

Truncated SVD filter	Tikhonov filter	Exponential filter
$\phi_i = \begin{cases} 1 & \text{if } \sigma_i > \tau \\ 0 & \text{if } \sigma_i \leq \tau \end{cases}$	$\phi_i = \dfrac{\sigma_i^2}{\sigma_i^2 + \alpha^2}$	$\phi_i = 1 - e^{-\sigma_i^2/\alpha^2}$

Note that using a Taylor series expansion of the exponential term in the exponential filter, it is not difficult to see that the Tikhonov filter is a truncated approximation of the exponential filter. Moreover, the Tikhonov filter has an equivalent variational form, which is described in section "Variational Regularization and Constraints".

Observe that each of the filtering methods has a parameter (e.g., in the above examples, τ and α) that needs to be chosen to specify how much filtering is done. Appropriate values depend on properties of the matrix $\mathbf{A}$ (i.e., on its singular values and singular vectors) as well as on the data, $\mathbf{b}$. Some techniques to help guide the choice of the regularization parameter are discussed in section "Choosing Regularization Parameters."

Because the SVD can be very expensive to compute for large matrices, this explicit filtering approach is generally not used for large-scale inverse problems. There are some exceptions, though, if $\mathbf{A}$ is highly structured. For example, suppose $\mathbf{A}$ can be decomposed as a Kronecker product,

$$\mathbf{A} = \mathbf{A}_r \otimes \mathbf{A}_c = \begin{bmatrix} a_{11}^{(r)}\mathbf{A}_c & a_{12}^{(r)}\mathbf{A}_c & \cdots & a_{1n}^{(r)}\mathbf{A}_c \\ a_{21}^{(r)}\mathbf{A}_c & a_{22}^{(r)}\mathbf{A}_c & \cdots & a_{2n}^{(r)}\mathbf{A}_c \\ \vdots & \vdots & & \vdots \\ a_{n1}^{(r)}\mathbf{A}_c & a_{n2}^{(r)}\mathbf{A}_c & \cdots & a_{nn}^{(r)}\mathbf{A}_c \end{bmatrix}$$

where $\mathbf{A}_c$ is an $m \times m$ matrix, and $\mathbf{A}_r$ is an $n \times n$ matrix with entries denoted by $a_{ij}^{(r)}$. Then this block structure can be exploited when computing the SVD and when implementing filtering algorithms [51].

It is also sometimes possible to use an alternative factorization. Specifically, suppose that

$$\mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^*,$$

where $\mathbf{\Lambda}$ is a diagonal matrix and $\mathbf{Q}^*$ is the complex conjugate transpose of $\mathbf{Q}$, with $\mathbf{Q}^*\mathbf{Q} = \mathbf{I}$. This is called a spectral factorization, where the columns of $\mathbf{Q}$ are eigenvectors and the diagonal elements of $\mathbf{\Lambda}$ are the eigenvalues of $\mathbf{A}$. Although every matrix has an SVD, only normal matrices (i.e., matrices that satisfy $\mathbf{A}^*\mathbf{A} = \mathbf{A}\mathbf{A}^*$) have a spectral decomposition. However, if $\mathbf{A}$ has a spectral factorization, then it can be used, in place of the SVD, to implement the filtering methods described in this section. The advantage is that it is sometimes more computationally convenient to compute a spectral decomposition than an SVD; an example of this is given in section "Linear Example: Deconvolution."

Variational Regularization and Constraints

Variational regularization methods have the form

$$\min_{\mathbf{x}} \left\{ \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2 + \alpha^2 \mathcal{J}(\mathbf{x}) \right\}, \tag{11}$$

where the regularization operator $\mathcal{J}$ and the regularization parameter α must be chosen. The variational form provides a lot of flexibility. For example, one could include additional constraints on the solution, such as nonnegativity, or it may be preferable to replace the least squares criterion with the Poisson log likelihood function [3–5]. As with filtering, there are many choices for the regularization operator, $\mathcal{J}$, such as Tikhonov, total variation [16, 85, 99], and sparsity constraints [13, 34, 94]:

Tikhonov	Total variation	Sparsity
$\mathcal{J}(\mathbf{x}) = \|\mathbf{L}\mathbf{x}\|_2^2$	$\mathcal{J}(\mathbf{x}) = \left\| \sqrt{(\mathbf{D}_h\mathbf{x})^2 + (\mathbf{D}_v\mathbf{x})^2} \right\|_1$	$\mathcal{J}(\mathbf{x}) = \|\|\mathbf{\Phi}\mathbf{x}\|\|_1$

Tikhonov regularization, which was first proposed and studied extensively in the early 1960s [69, 83, 89–91], is perhaps the most well-known approach to regularizing ill-posed problems. $\mathbf{L}$ is typically chosen to be the identity matrix, or a discrete approximation to a derivative operator, such as the Laplacian. If $\mathbf{L} = \mathbf{I}$, then it is not difficult to show that the resulting variational form of Tikhonov regularization, namely,

$$\min_{\mathbf{x}} \left\{ \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2 + \alpha^2 \|\mathbf{x}\|_2^2 \right\}, \tag{12}$$

can be written in an equivalent filtering framework by replacing $\mathbf{A}$ with its SVD [49].

For total variation, $\mathbf{D}_h$ and $\mathbf{D}_v$ denote discrete approximations of horizontal and vertical derivatives of the 2D image $\mathbf{x}$, and the approach extends to 3D images in an obvious way. Efficient and stable implementation of total variation regularization is a nontrivial problem; see [16, 99] and the references therein for further details.

In the case of sparse reconstructions, the matrix $\mathbf{\Phi}$ represents a basis in which the image, $\mathbf{x}$, is sparse. For example, for astronomical images that contain a few bright objects surrounded by a significant amount of black background, an appropriate choice for $\mathbf{\Phi}$ might be the identity matrix. Clearly, the choice of $\mathbf{\Phi}$ is highly dependent on the structure of the image $\mathbf{x}$. The usage of sparsity constraints for regularization is currently a very active field of research, with many open problems. We refer interested readers to the chapter in this handbook on *compressive sensing*, and the references therein.

We also mention that when the majority of the elements in the image $\mathbf{x}$ are zero or near zero, as may be the case for astronomical or medical images, it may be wise to enforce nonnegativity constraints on the solution [4, 5, 99]. This requires that each element of the computed solution $\mathbf{x}$ is not negative, which is often written as $\mathbf{x} \geq 0$. Though these constraints add a level of difficulty when solving, they can produce results that are more feasible than when nonnegativity is ignored.

Finally, it should be noted that depending on the structure of matrix $\mathbf{A}$, the type of regularization, and the additional constraints, a variety of optimization algorithms can be used to solve (11). In some cases, it is possible to use a very efficient filtering approach, but typically it is necessary to use an iterative method.

Iterative Regularization

As mentioned in section "Variational Regularization and Constraints," iterative methods are often needed to solve the variational form of the regularized problem. An alternate approach to using variational regularization is to simply apply the iterative method to the least squares problem,

$$\min_{\mathbf{x}} \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2 .$$

Note that if an iterative method applied to this unregularized problem is allowed to "converge," it will converge to an inverse solution, $\mathbf{x}_{\text{inv}}$, which is corrupted by noise (recall the discussion in section "SVD Analysis"). However, many iterative methods have the property (provided the problem on which it is applied satisfies the discrete Picard condition) that the early iterations reconstruct components of the solution corresponding to large singular values, while components corresponding to small singular values are reconstructed at later iterations. Thus, there is an observed "semi-convergence" behavior in the quality of the reconstruction, whereby the approximate solution improves at early iterations and then degrades at later iterations (a more detailed discussion of this behavior is given in section "Hybrid Iterative-Direct Regularization" in the context of the iterative method LSQR). If

the iteration is terminated at an appropriate point, a regularized approximation of the solution is computed. Thus, the iteration index acts as the regularization parameter, and the associated scheme is referred to as an *iterative regularization method*.

Many algorithms can be used as iterative regularization methods, including Landweber [61], steepest descent, and the conjugate gradient method (e.g., for nonsymmetric problems, the CGLS implementation [8] or the LSQR implementation [80, 81], and for symmetric indefinite problems, the MR-II implementation [43]). Most iterative regularization methods can be put into a general framework associated with solving the minimization problem

$$\min f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{A}^T\mathbf{A}\mathbf{x} - \mathbf{x}^T\mathbf{A}^T\mathbf{b} \tag{13}$$

with a general iterative method of the form

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \rho_k\mathbf{M}_k(\mathbf{A}^T\mathbf{b} - \mathbf{A}^T\mathbf{A}\mathbf{x}_k) = \mathbf{x}_k + \rho_k\mathbf{M}_k\mathbf{r}_k, \tag{14}$$

where $\mathbf{r}_k = \mathbf{A}^T\mathbf{b} - \mathbf{A}^T\mathbf{A}\mathbf{x}_k$. With specific choices of ρ_k and $\mathbf{M}_k$, one can obtain a variety of well-known iterative methods:

- The Landweber method is obtained by taking $\rho_k = \rho$ (i.e., ρ remains constant for each iteration) and $\mathbf{M}_k = \mathbf{I}$ (the identity matrix). Due to its very slow convergence, this classic approach is not often used for linear inverse problems. However, it is very easy to analyze the regularization properties of the Landweber iteration, and it can be useful for certain large-scale nonlinear ill-posed inverse problems.
- The steepest descent method is produced if $\mathbf{M}_k = \mathbf{I}$ is again fixed as the identity, but now ρ_k is chosen to minimize the residual at each iteration. That is, ρ_k is chosen as

$$\rho_k = \arg\min_{\rho>0} f(\mathbf{x}_k + \rho\mathbf{r}_k).$$

- Again, this method typically has very slow convergence, but with proper preconditioning it may be competitive with other methods.
- It is also possible to obtain the conjugate gradient method by setting $\mathbf{M}_0 = \mathbf{I}$ and $\mathbf{M}_{k+1} = \mathbf{I} - \frac{\mathbf{s}_k\mathbf{y}_k^T}{\mathbf{y}_k^T\mathbf{s}_k}$, where $\mathbf{s}_k = \mathbf{x}_{k+1} - \mathbf{x}_k$ and $\mathbf{y}_k = \mathbf{A}^T\mathbf{A}(\mathbf{x}_{k+1} - \mathbf{x}_k)$. As with the steepest descent method, ρ_k is chosen to minimize the residual at each iteration. Generally, the conjugate gradient method converges much more quickly than Landweber or steepest descent.

Other iterative algorithms that can be put into this general frame work include the Brakhage ν methods [10] and Barzilai and Borwein's lagged steepest descent scheme [6].

Hybrid Iterative-Direct Regularization

One of the main disadvantages of iterative regularization methods is that it can be very difficult to determine appropriate stopping criteria. To address this problem, work has been done to develop hybrid methods that combine variational approaches with iterative methods. That is, an iterative method, such as the LSQR implementation of the conjugate gradient method, is applied to the least squares problem $\min_{\mathbf{x}} ||\mathbf{A}\mathbf{x} - \mathbf{b}||_2^2$, and variational regularization is incorporated within the iteration process. To understand how this can be done, it is necessary to briefly describe how the LSQR iterates are computed.

LSQR is based on the Golub-Kahan (sometimes referred to as Lanczos) bidiagonalization (GKB) process. Given an $m \times n$ matrix $\mathbf{A}$ and vector $\mathbf{b}$, the kth GKB iteration computes an $m \times (k+1)$ matrix $\mathbf{W}_k$, an $n \times k$ matrix $\mathbf{Y}_k$, an $n \times 1$ vector $\mathbf{y}_{k+1}$, and a $(k+1) \times k$ bidiagonal matrix $\mathbf{B}_k$ such that

$$\mathbf{A}^T \mathbf{W}_k = \mathbf{Y}_k \mathbf{B}_k^T + \gamma_{k+1} \mathbf{y}_{k+1} \mathbf{e}_{k+1}^T \tag{15}$$

$$\mathbf{A}\mathbf{Y}_k = \mathbf{W}_k \mathbf{B}_k, \tag{16}$$

where $\mathbf{e}_{k+1}$ denotes the $(k+1)$st standard unit vector and $\mathbf{B}_k$ has the form

$$\mathbf{B}_k = \begin{bmatrix} \gamma_1 & & & \\ \beta_2 & \gamma_2 & & \\ & \ddots & \ddots & \\ & & \beta_k & \gamma_k \\ & & & \beta_{k+1} \end{bmatrix}. \tag{17}$$

Matrices $\mathbf{W}_k$ and $\mathbf{Y}_k$ have orthonormal columns, and the first column of $\mathbf{W}_k$ is $\mathbf{b}/||\mathbf{b}||_2$. Given these relations, an approximate solution $\mathbf{x}_k$ can be computed from the *projected* least squares problem

$$\min_{\mathbf{x} \in R(\mathbf{Y}_k)} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 = \min_{\hat{\mathbf{x}}} \|\mathbf{B}_k \hat{\mathbf{x}} - \beta \mathbf{e}_1\|_2^2 \tag{18}$$

where $\beta = ||\mathbf{b}||_2$ and $\mathbf{x}_k = \mathbf{Y}_k \hat{\mathbf{x}}$. An efficient implementation of LSQR does not require storing the matrices $\mathbf{W}_k$ and $\mathbf{Y}_k$ and uses an efficient updating scheme to compute $\hat{\mathbf{x}}$ at each iteration; see [81] for details.

An important property of GKB is that for small values of k, the singular values of the matrix $\mathbf{B}_k$ approximate very well certain singular values of $\mathbf{A}$, with the quality of the approximation depending on the relative spread of the singular values; specifically, the larger the relative spread, the better the approximation [8, 37, 87]. For ill-posed inverse problems, the singular values decay to and cluster at zero, such as $\sigma_i = O(i^{-c})$ where $c > 1$ or $\sigma_i = O(c^i)$ where $0 < c < 1$ and $i = 1, 2, \ldots, n$ [95, 96]. Thus, the relative gap between large singular values is generally much larger than the relative gap between small singular values. Therefore, if the GKB

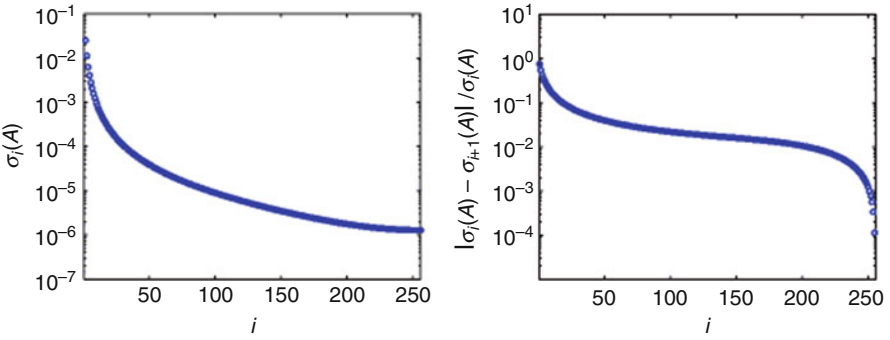


Fig. 2 This figure shows plots of the singular values of $\mathbf{A}$, denoted as $\sigma_i(\mathbf{A})$ (*left plot*), and the relative spread of $\mathbf{A}$'s singular values (*right plot*)

iteration is applied to a linear system arising from discretization of an ill-posed inverse problem, then the singular values of $\mathbf{B}_k$ converge very quickly to the largest singular values of $\mathbf{A}$. The following example illustrates this situation.

Example 1. Consider a linear system obtained by discretization of a one-dimensional first-kind Fredholm integral equation of the form (3), where the kernel $k(s,t)$ is given by the Green's function for the second derivative and which is constructed using *deriv2* in the MATLAB package *Regularization Tools* [48]. Although this is not an imaging example, it is a small-scale canonical ill-posed inverse problem that has properties found in imaging applications. The *deriv2* function constructs an $n \times n$ matrix $\mathbf{A}$ from the kernel

$$k(s,t) = \begin{cases} s(t-1) & \text{if } s < t \\ t(s-1) & \text{if } s \geq t \end{cases}$$

defined on $[0,\ 1] \times [0,\ 1]$. We use $n = 256$. There are also several choices for constructing vectors $\mathbf{x}_{\text{exact}}$ and $\mathbf{b}_{\text{exact}}$ (see [48]), but we focus only on the matrix $\mathbf{A}$ in this example.

Figure 2 shows a plot of the singular values of $\mathbf{A}$ and their relative spread; that is,

$$\frac{\sigma_i(\mathbf{A}) - \sigma_{i+1}(\mathbf{A})}{\sigma_i(\mathbf{A})},$$

where the notation $\sigma_i(\mathbf{A})$ is used to denote the ith largest singular value of $\mathbf{A}$. Figure 2 clearly illustrates the properties of ill-posed inverse problems; the singular values of $\mathbf{A}$ decay to and cluster at 0. Moreover, it can be observed that in general the relative gap of the singular values is larger for the large singular values and smaller for the smaller singular values. Thus, for small values of k, the singular values of $\mathbf{B}_k$ converge quickly to the large singular values of $\mathbf{A}$. This can be seen in Fig. 3,

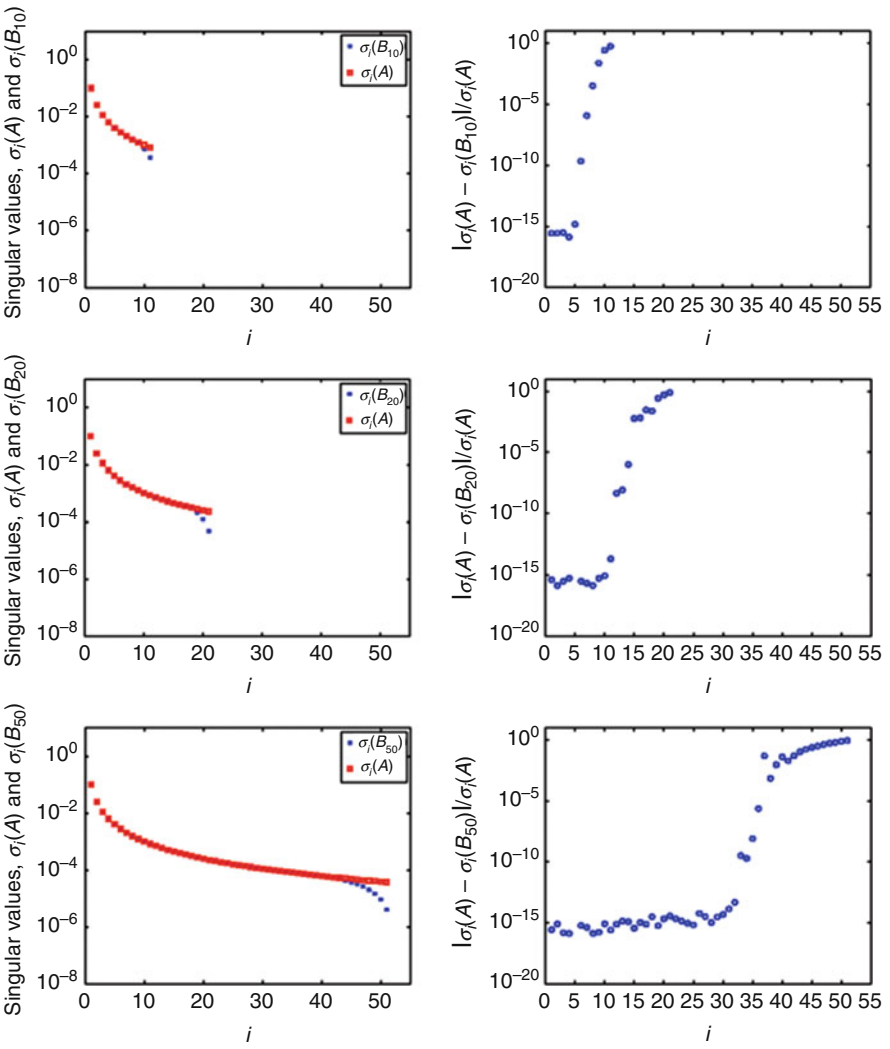


Fig. 3 The plots in the *left column* of this figure show the singular values of $\mathbf{A}$, denoted as $\sigma_i(\mathbf{A})$, along with the singular values of $\mathbf{B}_k$, denoted as $\sigma_i(\mathbf{B}_k)$, for $k = 10$, 20, and 50. The plots in the *right column* show the relative difference, $\frac{|\sigma_i(\mathbf{A}) - \sigma_i(\mathbf{B}_k)|}{\sigma_i(\mathbf{A})}$

which compare the singular values of $\mathbf{A}$ with those of the bidiagonal matrix $\mathbf{B}_k$ for $k = 10, 20$, and 50.

This example implies that if LSQR is applied to the least squares problem $\min_{\mathbf{x}} ||\mathbf{A}\mathbf{x} - \mathbf{b}||_2$, then at early iterations the approximate solutions $\mathbf{x}_k$ will be in a subspace that approximates a subspace spanned by the large singular components of $\mathbf{A}$. Thus, for $k < n, \mathbf{x}_k$ is a regularized solution. However, eventually $\mathbf{x}_k$

should converge to the inverse solution, which is corrupted with noise (recall the discussion in section "Iterative Regularization"). This means that the iteration index k plays the role of a regularization parameter; if k is too small, then the computed approximation $\mathbf{x}_k$ is an over-smoothed solution, while if k is too large, $\mathbf{x}_k$ is corrupted with noise. Again, we emphasize that this semi-convergence behavior requires that the problem satisfies the discrete Picard condition. More extensive theoretical arguments of this semi-convergence behavior of conjugate gradient methods can be found elsewhere; see [43] and the references therein.

Instead of early termination of the iteration, hybrid approaches enforce regularization at each iteration of the GKB method. Hybrid methods were first proposed by O'Leary and Simmons in 1981 [78] and later by Björck in 1988 [7]. The basic idea is to regularize the projected least squares problem (18) involving $\mathbf{B}_k$, which can be done very cheaply because of the smaller size of $\mathbf{B}_k$. More specifically, because the singular values of $\mathbf{B}_k$ approximate those of $\mathbf{A}$, as the GKB iteration proceeds, the matrix $\mathbf{B}_k$ becomes more ill-conditioned. The iteration can be stabilized by including Tikhonov regularization in the projected least square problem (18) to obtain

$$\min_{\hat{\mathbf{x}}} \left\{ \|\mathbf{B}_k \hat{\mathbf{x}} - \beta \mathbf{e}_1\|_2^2 + \alpha^2 \|\hat{\mathbf{x}}\|_2^2 \right\} \tag{19}$$

where again $\beta = ||\mathbf{b}||_2$ and $\mathbf{x}_k = \mathbf{Y}_k \hat{\mathbf{x}}$. Thus, at each iteration it is necessary to solve a regularized least squares problem involving a bidiagonal matrix $\mathbf{B}_k$. Notice that since the dimension of $\mathbf{B}_k$ is very small compared to $\mathbf{A}$, it is much easier to solve for $\hat{\mathbf{x}}$ in Eq. (19) than it is to solve for $\mathbf{x}$ in the full Tikhonov regularized problem (12). More importantly, when solving Eq. (19) one can use sophisticated parameter choice methods to find a suitable α at each iteration.

To summarize, hybrid methods have the following benefits:

- Powerful regularization parameter choice methods can be implemented efficiently on the projected problem.
- Semi-convergence behavior of the relative errors observed in LSQR is avoided, so an imprecise (over) estimate of the stopping iteration does not have a deleterious effect on the computed solution.

Realizing these benefits in practice, though, is nontrivial. Thus, various authors have considered computational and implementation issues, such as robust approaches to choose regularization parameters and stopping iterations; see, for example, [9, 11, 21, 45, 60, 62, 78]. We also remark that our discussion of hybrid methods focused on the case of Tikhonov regularization with $\mathbf{L} = \mathbf{I}$. Implementation of hybrid methods when $\mathbf{L}$ is not the identity matrix, such as a differentiation operator, can be nontrivial; see, for example, [50, 59].

Choosing Regularization Parameters

Each of the regularization methods discussed in this section requires choosing a *regularization parameter*. It is a nontrivial matter to choose "optimal" regularization parameters, but there are methods that can be used as guides. Some require a priori information, such as a bound on the noise or a bound on the solution. Others attempt to estimate an appropriate regularization parameter directly from the given data.

To describe some of the more popular parameter choice methods, let $\mathbf{x}_{\text{reg}}$ denote a solution computed by a particular regularization method:

- **Discrepancy Principle.** In this approach, a solution is sought such that

$$\left\|\mathbf{b} - \mathbf{A}\mathbf{x}_{\text{reg}}\right\|_2 = \tau \left\|\boldsymbol{\eta}\right\|_2$$

  Where $\tau > 1$ is a predetermined number [71]. This is perhaps the easiest of the methods to implement, and there are substantial theoretical results establishing its behavior in the presence of noise. However, it is necessary to have a good estimate for $||\boldsymbol{\eta}||_2$.
- **Generalized Cross Validation.** The idea behind generalized cross validation (GCV) is that if one data point is removed from the problem, then a good regularized solution should predict that missing data point well. If α is the regularization parameter used to obtain $\mathbf{x}_{\text{reg}}$, then it can be shown [36] that the GCV method chooses α to minimize the function

$$G(\alpha) = \frac{\left\|\mathbf{b} - \mathbf{A}\mathbf{x}_{\text{reg}}\right\|^2}{\left(\text{trace}\left(\mathbf{I} - \mathbf{A}\mathbf{A}^{\dagger}_{\text{reg}}\right)\right)^2}.$$

  Where $\mathbf{A}^{\dagger}_{\text{reg}}$ is the matrix such that $\mathbf{x}_{\text{reg}} = \mathbf{A}^{\dagger}_{\text{reg}}\mathbf{b}$. For example, in the case of Tikhonov regularization (12),

$$\mathbf{A}^{\dagger}_{\text{reg}} = (\mathbf{A}^T\mathbf{A} + \alpha^2\mathbf{I})^{-1}\mathbf{A}^T.$$

  A weighted version of GCV, W-GCV, finds a regularization parameter to minimize

$$G_{\omega}(\alpha) = \frac{\left\|\mathbf{b} - \mathbf{A}\mathbf{x}_{\text{reg}}\right\|^2}{\left(\text{trace}\left(\mathbf{I} - \omega\mathbf{A}\mathbf{A}^{\dagger}_{\text{reg}}\right)\right)^2}.$$

  W-GCV is sometimes more effective at choosing regularization parameters than the standard GCV function for certain classes of problems. Setting the weight $\omega = 1$ gives the standard GCV method, while $\omega < 1$ produces less smooth solutions and $\omega > 1$ produces smoother solutions. Further details about W-GCV can be found in [21].

- **L-Curve.** This approach attempts to balance the size of the discrepancy (i.e., residual) produced by the regularized solution with the size of the solution. In the context of Tikhonov regularization, this can often be found by a log-log scale plot of $||\mathbf{b} - \mathbf{A}\mathbf{x}_{\text{reg}}||_2$ versus $||\mathbf{x}_{\text{reg}}||_2$ for all possible regularization parameters. This plot often produces an L-shaped curve, and the solution corresponding to the corner of the L indicates a good balance between discrepancy and size of the solution. This observation was first made by Lawson and Hanson [63] and later studied extensively, including efficient numerical schemes to find the corner of the L (i.e., the point of maximum curvature) by Hansen [46, 52]. Although the L-curve tends to work well for many problems, some concerns about its effectiveness have been reported in the literature; see [44, 98].

There exist many other parameter choice methods besides the ones discussed above; for more information, see [28, 49, 99] and the references therein.

A proper choice of the regularization parameter is critical. If the parameter is chosen too small, then too much noise will be introduced in the computed solution. On the other hand, if the parameter is too large, the regularized solution may become over smoothed and may not contain as much information about the true solution as it could. However, it is important to keep in mind that no parameter choice method is "fool proof," and it may be necessary to solve the problem with a variety of parameters and to use knowledge of the application to help decide which solution is best.

Separable Inverse Problems

Separable nonlinear inverse problems,

$$\mathbf{b} = \mathbf{A}\left(\mathbf{x}_{\text{exact}}^{(n\ell)}\right)\mathbf{x}_{\text{exact}}^{(\ell)} + \boldsymbol{\eta}, \tag{20}$$

arise in many imaging applications, such as blind deconvolution (see section "Multi-frame Blind Deconvolution"), super resolution (which is an example of image data fusion) [18, 27, 57, 76], the reconstruction of 3D macromolecular structures from 2D electron microscopy images of cryogenically frozen samples (Cryo-EM) [22, 35, 55, 66, 82, 88], and seismic imaging applications [40]. One could consider Eq. (20) as a general nonlinear inverse problem and use the approaches discussed in section "Nonlinear Inverse Problems" to compute regularized solutions. However, this section considers approaches that exploit the separability of the problem. In particular, some of the regularization methods described in section "Linear Problems," such as variational and iterative regularization, can be adapted to Eq. (20). To illustrate, consider the general Tikhonov regularized least squares problem:

$$\min_{\mathbf{x}^{(\ell)},\mathbf{x}^{(n\ell)}} \left\{ \|\mathbf{A}\left(\mathbf{x}^{(n\ell)}\right)\mathbf{x}^{(\ell)} - \mathbf{b}\|_2^2 + \alpha^2 \|\mathbf{x}^{(\ell)}\|_2^2 \right\} = \min_{\mathbf{x}^{(\ell)},\mathbf{x}^{(n\ell)}} \left\| \begin{bmatrix} \mathbf{A}(\mathbf{x}^{(n\ell)}) \\ \alpha\mathbf{I} \end{bmatrix} \mathbf{x}^{(\ell)} - \begin{bmatrix} \mathbf{b} \\ \mathbf{0} \end{bmatrix} \right\|_2^2. \tag{21}$$

Three approaches to solve this nonlinear least squares problem are outlined in this section.

Fully Coupled Problem

The nonlinear least squares problem given in Eq. (21) can be rewritten as

$$\min_{\mathbf{x}} \phi(\mathbf{x}) = \min_{\mathbf{x}} \frac{1}{2} \|\boldsymbol{\rho}(\mathbf{x})\|_2^2, \tag{22}$$

where

$$\boldsymbol{\rho}(\mathbf{x}) = \boldsymbol{\rho}\left(\mathbf{x}^{(\ell)}, \mathbf{x}^{(n\ell)}\right) = \begin{bmatrix} \mathbf{A}\left(\mathbf{x}^{(n\ell)}\right) \\ \alpha\mathbf{I} \end{bmatrix} \mathbf{x}^{(\ell)} - \begin{bmatrix} \mathbf{b} \\ \mathbf{0} \end{bmatrix}, \quad \text{and} \quad \mathbf{x} = \begin{bmatrix} \mathbf{x}^{(\ell)} \\ \mathbf{x}^{(n\ell)} \end{bmatrix}$$

Nonlinear least squares problems are solved iteratively, with algorithms having the general form:

Algorithm 1: General Iterative Algorithm

1 Choose initial $\mathbf{x}_0 = \begin{bmatrix} \mathbf{x}_0^{(\ell)} \\ \mathbf{x}_0^{(n\ell)} \end{bmatrix}$

2 for $k = 0, 1, 2, \ldots$
- Choose a step direction, $\mathbf{d}_k$
- Determine step length, τ_k
- Update the solution: $\mathbf{x}_{k+1} = \mathbf{x}_k + \tau_k \mathbf{d}_k$
- Stop when a minimum of the objective is obtained

end

Typically, $\mathbf{d}_k$ is chosen to approximate the Newton direction,

$$\mathbf{d}_k = -\left(\hat{\phi}''(\mathbf{x}_k)\right)^{-1} \phi'(\mathbf{x}_k),$$

where $\hat{\phi}''$ is an approximation of ϕ'', $\phi' = \mathbf{J}_\phi^T \boldsymbol{\rho}$ and $\mathbf{J}_\phi$ is the Jacobian matrix

$$\mathbf{J}_\phi = \begin{bmatrix} \frac{\partial \boldsymbol{\rho}\left(\mathbf{x}^{(\ell)}, \mathbf{x}^{(n\ell)}\right)}{\partial \mathbf{x}^{(\ell)}} & \frac{\partial \boldsymbol{\rho}\left(\mathbf{x}^{(\ell)}, \mathbf{x}^{(n\ell)}\right)}{\partial \mathbf{x}^{(n\ell)}} \end{bmatrix}.$$

In the case of the Gauss-Newton method, which is often recommended for nonlinear least squares problems, $\hat{\phi}'' = \mathbf{J}_\phi^T \mathbf{J}_\phi$.

This general Gauss-Newton approach can work well, but constructing and solving the linear systems required to update $\mathbf{d}_k$ can be very expensive. Note that the dimension of the matrix $\mathbf{J}_\phi$ corresponds to the number of pixels in the image, $\mathbf{x}^{(\ell)}$, plus the number of parameters in $\mathbf{x}^{(n\ell)}$, and thus $\mathbf{J}_\phi$ may be on the order of $10^6 \times 10^6$. Thus, instead of using Gauss-Newton, it might be preferable to use a low-storage scheme such as the (nonlinear) conjugate gradient method. But there is a trade off – although the cost per iteration is reduced, the number of iterations needed to attain a minimum can increase significantly.

Relatively little research has been done on understanding and solving the fully coupled problem. For example, methods are needed for choosing regularization parameters. In addition, the rate of convergence of the linear and nonlinear terms may be quite different, and the effect this has on overall convergence rate is not well understood.

Decoupled Problem

Probably the simplest idea to solve the nonlinear least squares problem is to decouple it into two problems, one involving $\mathbf{x}^{(\ell)}$ and the other involving $\mathbf{x}^{(n\ell)}$. Specifically, the approach would have the form:

Algorithm 2: Block Coordinate Descent Iterative Algorithm

1 Choose initial $\mathbf{x}_0^{(n\ell)}$

2 for $k = 0, 1, 2, \ldots$

- Choose α_k and solve the linear problem:

$$\mathbf{x}_k^{(\ell)} = \arg\min_{\mathbf{x}^{(\ell)}} \left\| \mathbf{A}\left(\mathbf{x}_k^{(n\ell)}\right)\mathbf{x}^{(\ell)} - \mathbf{b} \right\|_2^2 + \alpha_k^2 \left\| \mathbf{x}^{(\ell)} \right\|_2^2$$

- Solve the nonlinear problem:

$$\mathbf{x}_{k+1}^{(n\ell)} = \arg\min_{\mathbf{x}^{(n\ell)}} \left\| \mathbf{A}\left(\mathbf{x}^{(n\ell)}\right)\mathbf{x}_k^{(\ell)} - \mathbf{b} \right\|_2^2 + \alpha_k^2 \left\| \mathbf{x}_k^{(\ell)} \right\|_2^2$$

- Stop when objectives are minimized

end

The advantage of this approach is that any of the approaches discussed in section "Linear Problems," including methods to determine α, can be used for the linear problem. The nonlinear problem involving $\mathbf{x}^{(n\ell)}$ requires using another iterative method, such as the Gauss-Newton method. However, there are often significantly fewer parameters than in the fully coupled approach discussed in the previous subsection. Thus, a Gauss-Newton method to update $\mathbf{x}_{k+1}^{(n\ell)}$ at each iteration is significantly more computationally tractable. A disadvantage to this approach, which is known in the optimization literature as block coordinate descent, is that it is not clear what are the practical convergence properties of the method. As

mentioned in the previous subsection, the rate of convergence of the linear and nonlinear terms may be quite different. Moreover, if the method does converge, it will typically be very slow (linear), especially for problems with tightly coupled variables [77].

Variable Projection Method

The variable projection method [38,39,58,79,86] exploits structure in the nonlinear least squares problem (21). The approach exploits the fact that $\phi(\mathbf{x}^{(\ell)}, \mathbf{x}^{(n\ell)})$ is linear in $\mathbf{x}^{(\ell)}$ and that $\mathbf{x}^{(n\ell)}$ contains relatively fewer parameters than $\mathbf{x}^{(\ell)}$. However, rather than explicitly separating variables $\mathbf{x}^{(\ell)}$ and $\mathbf{x}^{(n\ell)}$ as in coordinate descent, variable projection implicitly eliminates the linear parameters $\mathbf{x}^{(\ell)}$, obtaining a reduced cost functional that depends only on $\mathbf{x}^{(n\ell)}$. Then, a Gauss-Newton method is used to solve the optimization problem associated with the reduced cost functional. Specifically, consider

$$\psi\left(\mathbf{x}^{(n\ell)}\right) \equiv \phi\left(\mathbf{x}^{(\ell)}\left(\mathbf{x}^{(n\ell)}\right), \mathbf{x}^{(\ell)}\right)$$

where $\mathbf{x}^{(\ell)}(\mathbf{x}^{(n\ell)})$ is a solution of

$$\min_{\mathbf{x}^{(\ell)}} \phi\left(\mathbf{x}^{(\ell)}, \mathbf{x}^{(n\ell)}\right) = \min_{\mathbf{x}^{(\ell)}} \left\| \begin{bmatrix} \mathbf{A}\left(\mathbf{x}^{(n\ell)}\right) \\ \alpha \mathbf{I} \end{bmatrix} x^{(\ell)} - \begin{bmatrix} \mathbf{b} \\ \mathbf{0} \end{bmatrix} \right\|_2^2. \tag{23}$$

To use the Gauss-Newton algorithm to minimize the reduced cost functional $\psi(\mathbf{x}^{(n\ell)})$, it is necessary to compute $\psi'(\mathbf{x}^{(n\ell)})$. Note that because $\mathbf{x}^{(\ell)}$ solves (23), it follows that $\frac{\partial \phi}{\partial \mathbf{x}^{(\ell)}} = 0$ and thus

$$\psi'(y) = \frac{d\mathbf{x}}{d\mathbf{y}} \frac{\partial \phi}{\partial \mathbf{x}^{(\ell)}} + \frac{\partial \phi}{\partial \mathbf{x}^{(n\ell)}} = \frac{\partial \phi}{\partial \mathbf{x}^{(n\ell)}} = \mathbf{J}_\psi^T \boldsymbol{\rho},$$

where the Jacobian of the reduced cost functional is given by

$$J_\psi = \frac{\partial\left(\mathbf{A}\left(\mathbf{x}^{(n\ell)}\right)\mathbf{x}^{(\ell)}\right)}{\partial \mathbf{x}^{(n\ell)}}.$$

Thus, a Gauss-Newton method applied to the reduced cost functional has the basic form:

Although computing $\mathbf{J}_\psi$ is nontrivial, it is often much more tractable than constructing $\mathbf{J}_\phi$. In addition, the problem of variable convergence rates for the two sets of parameters, $\mathbf{x}^{(\ell)}$ and $\mathbf{x}^{(n\ell)}$, has been eliminated. Another big advantage of the variable projection method for large-scale inverse problems is that standard approaches, such as those discussed in section "Linear Problems," can be used to solve the linear regularized least squares problem at each iteration, including the schemes for estimating regularization parameters.

Algorithm 3: Variable Projection Gauss-Newton Algorithm

1 Choose initial $\mathbf{x}_0^{(n\ell)}$
2 for $k = 0, 1, 2, \ldots$
3 Choose α_k
4 $\mathbf{x}_k^{(\ell)} = \arg\min_{\mathbf{x}^{(\ell)}} \left\| \begin{bmatrix} \mathbf{A}\left(\mathbf{x}_k^{(n\ell)}\right) \\ \alpha_k \mathbf{I} \end{bmatrix} \mathbf{x}^{(\ell)} - \begin{bmatrix} \mathbf{b} \\ \mathbf{0} \end{bmatrix} \right\|_2$
5 $\mathbf{r}_k = \mathbf{b} - \mathbf{A}\left(\mathbf{x}_k^{(n\ell)}\right) \mathbf{x}_k^{(\ell)}$
6 $\mathbf{d}_k = \arg\min_d ||\mathbf{J}_\psi \mathbf{d} - \mathbf{r}_k||_2$
7 Determine step length τ_k
8 $\mathbf{x}_{k+1}^{(n\ell)} = \mathbf{x}_k^{(n\ell)} + \tau_k \mathbf{d}_k$
9 end

Nonlinear Inverse Problems

Developing regularization approaches for general nonlinear inverse problems can be significantly more challenging than the linear and separable nonlinear case. Theoretical tools such as the SVD that are used to analyze ill-posedness in the linear case are not available here, and previous efforts to extend these tools to the nonlinear case do not always apply. For example, a spectral analysis of the linearization of a nonlinear problem does not necessarily determine the degree of ill-posedness for the nonlinear problem [30]. Furthermore, convergence properties for nonlinear optimization require very strict assumptions that are often not realizable in real applications [28,29]. Nevertheless, nonlinear inverse problems arise in many important applications, motivating research on regularization schemes and general computational approaches. This section discusses some of this work.

One approach for nonlinear problems of the form

$$F(\mathbf{x}) = \mathbf{b} \tag{24}$$

is to reformulate the problem to find a zero of $F(\mathbf{x}) - \mathbf{b} = 0$. Then a Newton-like method, where the nonlinear function is repeatedly linearized around the current estimate, can be written as

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \rho_k \mathbf{p}_k \tag{25}$$

where $\mathbf{p}_k$ solves the Jacobian system

$$\mathbf{J}(\mathbf{x}_k)\mathbf{p} = \mathbf{b} - F(\mathbf{x}_k). \tag{26}$$

Though generally not symmetric, matrix and matrix-transpose multiplications with the Jacobian, whose elements are the first derivatives of $F(\mathbf{x})$, are typically computable. However, the main disadvantages of using this approach are that the

existence and uniqueness of a solution are not guaranteed and the sensitivity of solutions depends on the conditioning of the Jacobian. Furthermore, there is no natural merit function that can be monitored to help select the step length, ρ_k.

Another approach to solve (24) is to incorporate prior assumptions regarding the statistical distribution of the model and maximize the corresponding likelihood function. For example, an additive Gaussian noise model assumption under certain conditions corresponds to solving the following nonlinear least squares problem:

$$\min_{\mathbf{x}} \tfrac{1}{2} \|\mathbf{b} - F(\mathbf{x})\|_2^2 . \tag{27}$$

Since this is a standard nonlinear optimization problem, any optimization algorithm such as a gradient descent or Newton approach can be used here. For problem (27), the gradient vector can be written as $\mathbf{g}(\mathbf{x}) = \mathbf{J}(\mathbf{x})^T(F(\mathbf{x}) - \mathbf{b})$, and Hessian matrix can be written as $\mathbf{H}(\mathbf{x}) = \mathbf{J}(\mathbf{x})^T\mathbf{J}(\mathbf{x}) + \mathbf{Z}(\mathbf{x})$, where $\mathbf{Z}(\mathbf{x})$ includes second derivatives of $F(\mathbf{x})$. The main advantage of this approach is that a variety of line search methods can be used. However, the potential disadvantages of this approach are that the derivatives may be too difficult to compute or that negative eigenvalues introduced in $\mathbf{Z}(\mathbf{x})$ may cause problems in optimization algorithms. Some algorithms for solving nonlinear optimization problems are direct extensions of the iterative methods described in section "Iterative Regularization." The nonlinear Landweber iteration can be written as

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{J}(\mathbf{x}_k)^T \left(\mathbf{b} - F(\mathbf{x}_k)\right), \tag{28}$$

which reduces to the standard Landweber iteration if $F(\mathbf{x})$ is linear, and it can be easily extended to other gradient descent methods such as the steepest descent approach. Newton and Newton-type methods are also viable options for nonlinear optimization, resulting in iterates (25) where $\mathbf{p}_k$ solves

$$\mathbf{H}(\mathbf{x}_k)\mathbf{p} = -\mathbf{g}(\mathbf{x}_k). \tag{29}$$

Oftentimes, an approximation of the Hessian is used. For example, the Gauss-Newton algorithm, which takes $\mathbf{H} \approx \mathbf{J}(\mathbf{x}_k)^T\mathbf{J}(\mathbf{x}_k)$, is a preferred choice for large-scale problems because it ensures positive semi-definiteness, but it is not advisable for large residual problems or highly nonlinear problems [40]. Additionally, nonlinear conjugate gradient, truncated-Newton, or quasi-Newton methods, such as LBFGS can be good alternatives if storage is a concern. It is important to remark that finding a global minimizer for a nonlinear optimization problem is in general very difficult, especially since convexity of the objective function is typically not guaranteed, as in the linear case. Thus, it is very likely that a descent algorithm may get stuck in one of many local minima solutions.

When dealing with ill-posed problems, the general approach to incorporate regularization is to couple an iterative approach with a stopping criteria such as the discrepancy principle to produce reasonable solutions. In addition, for Newton-type

methods it is common to incorporate additional regularization for the inner system since the Jacobian or Hessian may become ill-conditioned. For example, including linear Tikhonov regularization in (26) would result in

$$\left(\mathbf{J}(\mathbf{x}_k)^T\mathbf{J}(\mathbf{x}_k) + \alpha^2\mathbf{I}\right)\mathbf{p} = \mathbf{J}(\mathbf{x}_k)^T(\mathbf{b} - F(\mathbf{x}_k)),$$

which is equivalent to a Levenberg-Marquardt iterate, where the update, $\mathbf{p}_k$, is the solution of a particular Tikhonov minimization problem:

$$\min_{\mathbf{p}} \|F(\mathbf{x}_k) + \mathbf{J}(\mathbf{x}_k)\mathbf{p} - \mathbf{b}\|_2^2 + \alpha^2 \|\mathbf{p}\|_2^2,$$

where $F(\mathbf{x})$ has been linearized around $\mathbf{x}_k$. Other variations for regularizing the update can be found in [28] and the references therein. Regularization for the inner system can also be achieved by solving the inner system inexactly using an iterative method and terminating the iterations early. These are called *inexact Newton* methods, and the early termination of the inner iterations is a good way not only to make this approach practical for large-scale problems but also to enforce regularization on the inner system.

The variational approaches discussed in section "Variational Regularization and Constraints" can be extended for the second class of algorithms where a likelihood function results in a nonlinear optimization problem. For example, after selecting a regularization operator $\mathcal{J}(\mathbf{x})$ and regularization parameter α for (27), the goal would be to solve a nonlinear optimization problem of the form

$$\min_{\mathbf{x}} \left\{ \|\mathbf{b} - F(\mathbf{x})\|_2^2 + \alpha^2 \mathcal{J}(\mathbf{x}) \right\}. \tag{30}$$

The flexibility in the choice of the regularization operator is nice, but selecting a good regularization parameter a priori can be a computationally demanding task, especially for large-scale problems. Some work on estimating the regularization parameter within a constrained optimization framework has been done [40, 41], but the most common approach for regularization of nonlinear ill-posed inverse problems is to use standard iterative methods to solve (27), where regularization is obtained via early termination of the iterations. It cannot be stressed enough that when using any iterative method to solve a nonlinear inverse problem where the regularization is not already incorporated, a good stopping iteration for the outer iteration that serves as a regularization parameter is imperative. See also [2, 28, 29, 32, 54, 92, 97] for additional references on nonlinear inverse problems.

4 Numerical Methods and Case Examples

Given a specific large-scale inverse problem from an imaging application, it can be nontrivial to implement the algorithms and regularization methods discussed in this chapter. Efficient computations require exploiting the structure of the problem.

Moreover, choosing specific regularization schemes and constraints requires knowledge about the physical process underlying the data collection process. A few illustrative examples, using the imaging applications described in section "Imaging Applications," are given in this section.

Linear Example: Deconvolution

Perhaps the most well-known and well-studied linear inverse problem is deconvolution. As discussed in section "Image Deblurring and Deconvolution," this spatially invariant image deblurring problem is modeled as

$$\mathbf{b} = \mathbf{A}\mathbf{x}_{\text{exact}} + \boldsymbol{\eta},$$

where $\mathbf{A}$ is a structured matrix that depends on the PSF and imposed boundary conditions. For example, if periodic boundary conditions are imposed on the blurring operation, then $\mathbf{A}$ has a circulant structure, and moreover, $\mathbf{A}$ has the spectral decomposition

$$\mathbf{A} = \mathbf{F}^* \boldsymbol{\Lambda} \mathbf{F},$$

where $\mathbf{F}$ is a matrix representing a d-dimensional discrete Fourier transform, which satisfies $\mathbf{F}^*\mathbf{F} = \mathbf{I}$. The matrix $\mathbf{F}$ does not need to be constructed explicitly. Instead, fast Fourier transform (FFT) functions can be used to implement matrix-vector multiplications with $\mathbf{F}$ and $\mathbf{F}^*$. Specifically, for 2D images:

$$\begin{aligned} \mathbf{F}\mathbf{x} &\Leftrightarrow \text{fft2(x)} \quad \text{(2D forward FFT)} \\ \mathbf{F}^*\mathbf{x} &\Leftrightarrow \text{ifft2(x)} \quad \text{(2D inverse FFT)} \end{aligned}$$

The main advantages are that FFT-based spectral filtering regularization algorithms are very easy to implement and extremely efficient; see [51] for implementation details.

To illustrate, consider the image data shown in Fig. 4, where the simulated observed image was obtained by convolving the PSF with the true image and adding 1 % Gaussian white noise. The PSF was constructed from a Gaussian blurring operator,

$$p_{ij} = \exp\left(\frac{-(i-k)^2 s_2^2 - (j-l)^2 s_1^2 + 2(i-k)(j-l)s_3^2}{2s_1^2 s_2^2 - 2s_3^4}\right) \tag{31}$$

centered at (k, l) (location of point source), with $s_1 = s_2 = 5$ and $s_3 = 0$. An FFT-based Tikhonov spectral filtering solution was computed, with regularization operator $\mathbf{L} = \mathbf{I}$ and regularization parameter $\alpha = 0.00544$, which was chosen using GCV. (All computations for this example were done with MATLAB. The

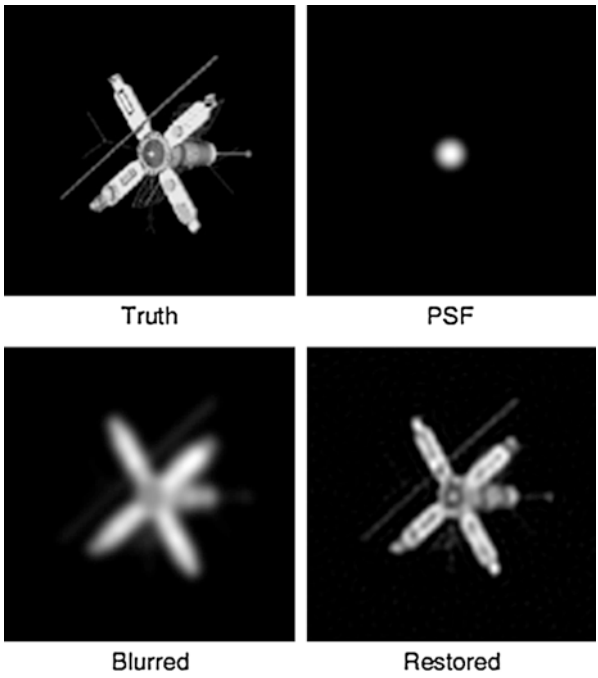


Fig. 4 Simulated data for an image deconvolution problem. The restored image was computed using an FFT-based spectral filtering method, with Tikhonov regularization and GCV-chosen regularization parameter

implementation of the FFT-based spectral filter used in this example is described in [51]. The MATLAB code, which is called *tik_fft.m*, can be found at http://www2.imm.dtu.dk/~pch/HNO/#.) The reconstructed image, which was computed in a fraction of a second on a standard laptop computer, is also shown in Fig. 4.

If there are significant details near the boundary of the image, then the periodic boundary condition assumption might not be an accurate representation of the details outside the viewable region. In this case, severe ringing artifacts can appear in the reconstructed image, and parameter choice methods may perform very poorly in these situations. Consider, for example, the image data shown in Fig. 5. The PSF is the same as in the previous example, but the blurred image contains features at the boundaries of the viewable region. The "restored" image in Fig. 5 was again computed using a Tikhonov spectral filtering solution with regularization operator $\mathbf{L} = \mathbf{I}$, and regularization parameter ($\alpha = 6.30 \times 10^{-5}$) was chosen using GCV. This noise-corrupted reconstructed image indicates that the regularization parameter chosen by GCV is too small.

It is possible that another parameter choice method would perform better, but it is also the case that imposing alternative boundary conditions may improve the situation. For example, reflective boundary conditions assume the image scene outside the viewable region is a mirror image of the details inside the viewable region. With this assumption, and if the PSF is also circularly symmetric, then the

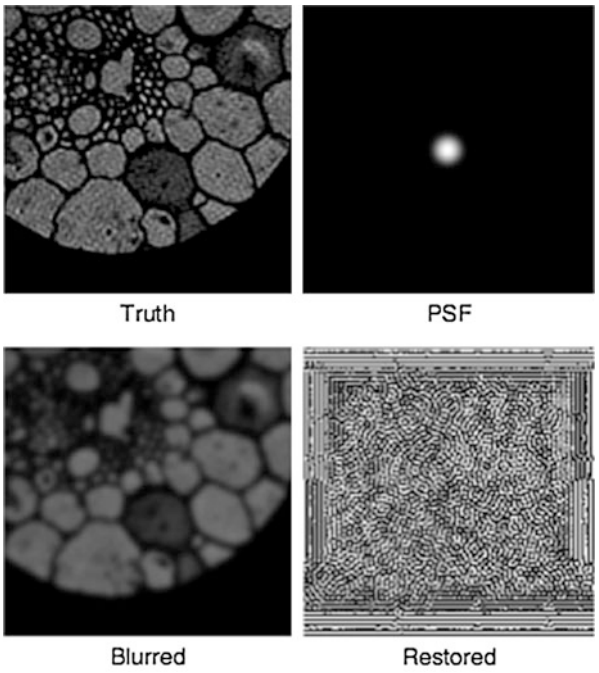


Fig. 5 Simulated data for an image deconvolution problem. The restored image was computed using an FFT-based spectral filtering method, with Tikhonov regularization and GCV-chosen regularization parameter

matrix $\mathbf{A}$ has a symmetric Toeplitz-plus-Hankel structure, and, moreover, $\mathbf{A}$ has the spectral decomposition

$$\mathbf{A} = \mathbf{C}^T \mathbf{\Lambda} \mathbf{C},$$

where $\mathbf{C}$ is a matrix representing a d-dimensional discrete cosine transform, which satisfies $\mathbf{C}^T\mathbf{C} = \mathbf{I}$. As with FFTs, the matrix $\mathbf{C}$ does not need to be constructed explicitly, and very efficient functions can be used to implement matrix-vector multiplications with $\mathbf{C}$ and $\mathbf{C}^T$, such as

$$\begin{aligned} \mathbf{C}\mathbf{x} &\Leftrightarrow \text{dct2(x)} \qquad \text{(2D forward DCT)} \\ \mathbf{C}^T\mathbf{x} &\Leftrightarrow \text{idct2(x)} \qquad \text{(2D inverse DCT)} \end{aligned}$$

In addition, DCT-based spectral filtering regularization algorithms are very easy to implement and are extremely efficient; see [51] for implementation details.

Figure 6 illustrates the superior performance that can be obtained if the boundary condition and the corresponding basis (in this case, DCT) is used in the spectral filtering deconvolution algorithms. Specifically, the image on the left in Fig. 6 was computed using a DCT-based Tikhonov spectral filtering method, with regularization operator $\mathbf{L} = \mathbf{I}$ and a GCV-chosen regularization parameter $\alpha = 4.83 \times 10^{-3}$. The image on the right was computed using the FFT-based Tikhonov filter, but instead of using the GCV-chosen regularization parameter (which produced the

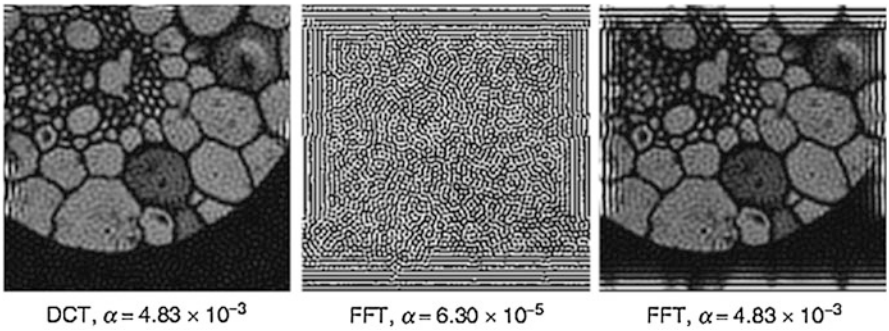


Fig. 6 These restored images were computed using DCT and FFT-based spectral filtering methods, with Tikhonov regularization. For the DCT and the middle FFT reconstructions, the regularization parameter α was chosen by GCV. The FFT reconstruction on the right was obtained using the same regularization parameter as was used for the DCT reconstruction

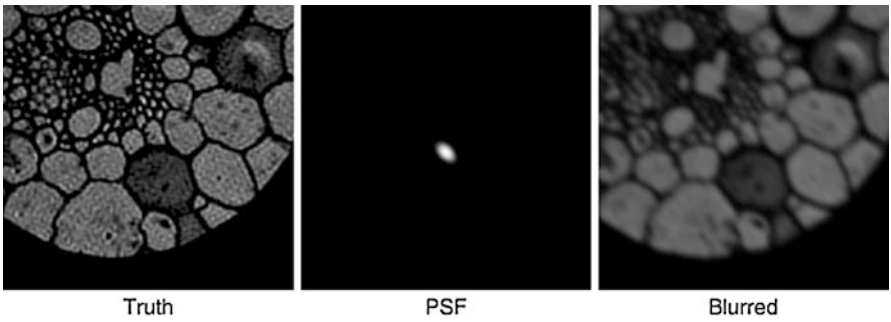


Fig. 7 Simulated deconvolution data with a nonsymmetric Gaussian PSF

poor reconstruction displayed in the middle of this figure), α was set to the same value used by the DCT reconstruction. This example clearly illustrates that the quality of the reconstruction, and the effectiveness of parameter choice methods, can depend greatly on the imposed boundary conditions and corresponding spectral basis. (As with previous examples, all computations described here were done with MATLAB. The implementation of the DCT-based spectral filter is described in [51]. The MATLAB code, which is called *tik_dct.m*, can be found at http://www2.imm.dtu.dk/~pch/HNO/.)

Spectral filtering methods work well for many deconvolution problems, but it may not always be possible to find a convenient basis that allows for efficient implementation. Consider, for example, the data shown in Fig. 7. The PSF in this figure was constructed using Eq. (31), with $s_1 = s_2 = 3$ and $s_3 = 2$, and results in a nonsymmetric PSF. As with the previous example, the FFT-based filter does not work well for this deconvolution problem because of its implicit assumption of

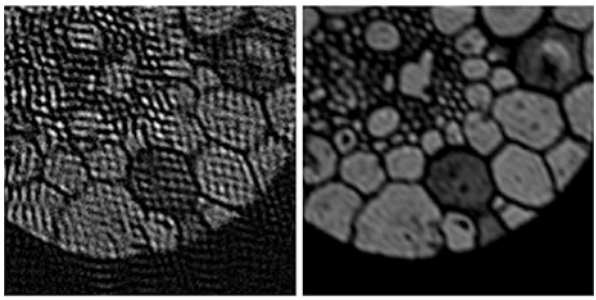


Fig. 8 These restored images were computed using a DCT-based spectral filtering method (*left*) and an iterative hybrid method (*right*)

periodic boundary conditions. Reflective boundary conditions are more appropriate, but the lack of circular symmetry in the PSF means that the DCT basis does not diagonalize the matrix $\mathbf{A}$. The reconstructed image on the left in Fig. 8 illustrates what happens if we attempt to reconstruct the image using a DCT-based Tikhonov filter.

An iterative method may be the best option for a problem such as this; one can impose any appropriate boundary condition (which only needs to be implemented in matrix-vector multiplications with $\mathbf{A}$ and $\mathbf{A}^T$) without needing to assume any symmetry or further structure in the PSF. The reconstructed image shown on the right in Fig. 8 was obtained using a hybrid approach described in section "Hybrid Iterative-Direct Regularization." Specifically, Tikhonov regularization is used for the projected subproblem, with regularization parameters chosen by W-GCV. The MATLAB software for this, which is called HyBR, is discussed in [21] and can be obtained from http://www.math.vt.edu/people/jmchung/hybr.html. For this particular example, HyBR terminated at iteration 21, with a regularization parameter $\alpha = 7.31 \times 10^{-2}$.

The examples in this subsection illustrate that many approaches can be used for the linear inverse problem deconvolution. It is possible that other methods, such as those that incorporate nonnegativity constraints, may produce better results than those presented here, but this is typical of all inverse problems. It would be impossible to give an exhaustive study and comparison in this chapter.

Separable Example: Multi-Frame Blind Deconvolution

In this section, multi-frame blind deconvolution (MFBD) is used to illustrate a numerical example of a separable (nonlinear) inverse problem,

$$\mathbf{b} = \mathbf{A}\left(\mathbf{x}_{\text{exact}}^{(n\ell)}\right)\mathbf{x}_{\text{exact}}^{(\ell)} + \boldsymbol{\eta}.$$

Recall from section "Multi-frame Blind Deconvolution" that in MFBD, a set of, say, m blurred images of an object are collected, and the aim is to simultaneously

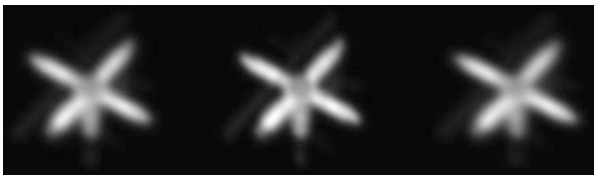


Fig. 9 Simulated MFBD data. The images were obtained by convolving the true satellite image from Fig. 4 with Gaussian PSFs using parameters given in Eq. (32) and then adding 1 % white noise

reconstruct an approximation of the true image as well as the PSFs (or the parameters that define the PSFs) associated with each of the observed blurred images. Such an example can be simulated using the Gaussian blurring kernel given in Eq. (31), and the true satellite image given in Fig. 4. Specifically, suppose using Eq. (31), three PSFs are constructed using the following values:

$$\mathbf{x}_{\text{exact}}^{(n\ell)} = \begin{bmatrix} 6.0516 \\ 5.8419 \\ 2.2319 \\ 5.4016 \\ 4.3802 \\ 2.1562 \\ 5.7347 \\ 6.8369 \\ 2.7385 \end{bmatrix} \begin{matrix} \left.\begin{matrix} \\ \\ \end{matrix}\right\} \text{Gaussian PSF parameters } s_1, s_2, s_3 \text{ for frame 1} \\ \left.\begin{matrix} \\ \\ \end{matrix}\right\} \text{Gaussian PSF parameters } s_1, s_2, s_3 \text{ for frame 2} \\ \left.\begin{matrix} \\ \\ \end{matrix}\right\} \text{Gaussian PSF parameters } s_1, s_2, s_3 \text{ for frame 3} \end{matrix} \tag{32}$$

Simulated observed image data can then be generated by convolving the PSFs constructed from these sets of parameters with the true satellite image and then adding 1 % white noise. The resulting simulated observed image frames are shown in Fig. 9.

Image reconstructions can then be computed using the variable projection Gauss-Newton algorithm described in section "Variable Projection Method." The Jacobian $\mathbf{J}_\psi$ can be constructed analytically for this problem (see, e.g., [19]), but a finite difference approach can also work very well. In the experiments reported here, centered differences were used to approximate the Jacobian.

The hybrid method implementation HyBR, described in the previous subsection, was used to choose α_k and to solve the linear subproblem for $\mathbf{x}_k^{(\ell)}$. The step length τ_k was chosen using an Armijo rule [77]. The initial guess for $\mathbf{x}_0^{(n\ell)}$ was

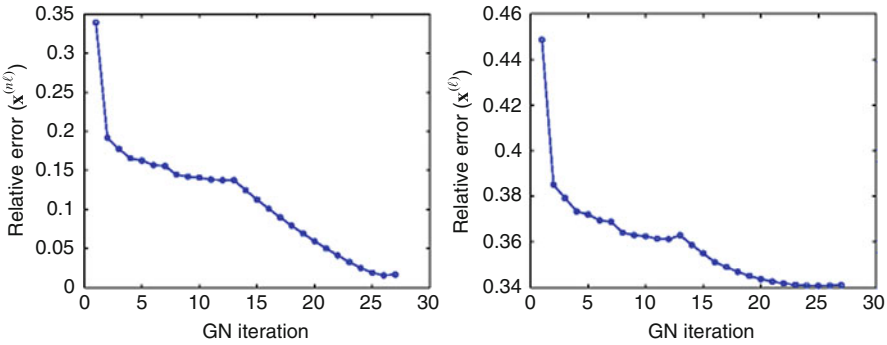


Fig. 10 Convergence results for MFBD. The relative error of the estimated PSF parameters at each iteration is shown in the *left plot*, while the relative error of the reconstructed image at each iteration is shown in the *right plot*

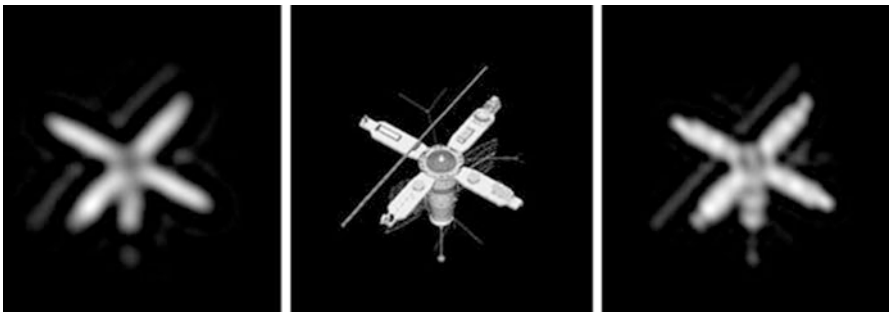


Fig. 11 On the *left* is the initial reconstructed image using $\mathbf{x}_0^{(n\ell)}$ and on the *right* is the final reconstructed image. The true image is displayed in the *middle* for comparison purposes

$$\mathbf{x}_0^{(n\ell)} = \begin{bmatrix} 7.0516 \\ 7.8369 \\ 3.2385 \\ 7.0516 \\ 7.8369 \\ 3.2385 \\ 7.0516 \\ 7.8369 \\ 3.2385 \end{bmatrix} \begin{matrix} \left.\vphantom{\begin{matrix}1\\1\\1\end{matrix}}\right\} \text{initial guess for } s_1, s_2, s_3 \text{ for frame 1} \\ \left.\vphantom{\begin{matrix}1\\1\\1\end{matrix}}\right\} \text{initial guess for } s_1, s_2, s_3 \text{ for frame 2} \\ \left.\vphantom{\begin{matrix}1\\1\\1\end{matrix}}\right\} \text{initial guess for } s_1, s_2, s_3 \text{ for frame 3} \end{matrix}$$

The results in Fig. 10 show the convergence behavior in terms of relative error at each iteration of the variable projection Gauss-Newton algorithm for this example. The left plot shows the convergence history of $\mathbf{x}_k^{(n\ell)}$, and the right plot shows the convergence history of $\mathbf{x}_k^{(\ell)}$. Note that the convergence behavior of both terms is very similar. Figure 11 shows the reconstructed image after the first variable projection

Gauss-Newton iteration (i.e., the initial reconstruction) and the reconstructed image after the last iteration of the algorithm.

Nonlinear Example: Tomosynthesis

Polyenergetic digital tomosynthesis is an example of a nonlinear inverse problem where the forward problem can be modeled as (8). The typical approach to compute an approximation of $\mathbf{x}_{\text{exact}}$ is to assume that the observed projection image is a realization of a Poisson random variable with mean values

$$\bar{\mathbf{b}}_\theta + \bar{\boldsymbol{\eta}} = \sum_{e=1}^{n_e} \varrho(e) \exp(-[s(e)\mathbf{A}_\theta \mathbf{x} + z(e)\mathbf{A}_\theta \mathbf{1}]) + \bar{\boldsymbol{\eta}}, \tag{33}$$

where $\bar{\boldsymbol{\eta}}$ is the mean of the additive noise. Then the maximum likelihood estimator (MLE) can be found by minimizing the negative log likelihood function:

$$-L_\theta(\mathbf{x}) = \sum_{i=1}^{M} \left(\bar{b}_\theta^{(i)} + \bar{\eta}^{(i)}\right) - b_\theta^{(i)} \log\left(\bar{b}_\theta^{(i)} + \bar{\eta}^{(i)}\right) + c, \tag{34}$$

where superscripts refer to entries in a vector and c is a constant term. A regularized estimate can be found by solving the following nonlinear optimization problem

$$\mathbf{x}_{\text{MLE}} = \min_{\mathbf{x}} \left\{ \sum_{\theta=1}^{n_\theta} -L_\theta(\mathbf{x}) \right\} \tag{35}$$

using a gradient descent or Newton-type algorithm and terminating the iterations before the noise enters the problem. For this example, the gradient of the objective function with respect to the 3D volume, $\mathbf{x}$, can be written as

$$g(\mathbf{x}_k) = \mathbf{A}^T \mathbf{v}_k$$

where the entries of vector $\mathbf{v}_k$ are given by

$$v^{(i)} = \left(\frac{b^{(i)}}{\bar{b}^{(i)} + \bar{\eta}^{(i)}} - 1 \right) \sum_{e=1}^{n_e} \varrho(e) s(e) \exp(-[s(e)\mathbf{a}_i^T \mathbf{x}_k + z(e)\mathbf{a}_i^T \mathbf{1}]).$$

The Hessian matrix can be written as

$$\mathbf{H}_k = \mathbf{A}^T \mathbf{W}_k \mathbf{A}$$

where $\mathbf{W}_k$ is a diagonal matrix with vector $\mathbf{w}_k$ on the diagonal. A mathematical formula for the values of the diagonal can be quite complicated as they depend on

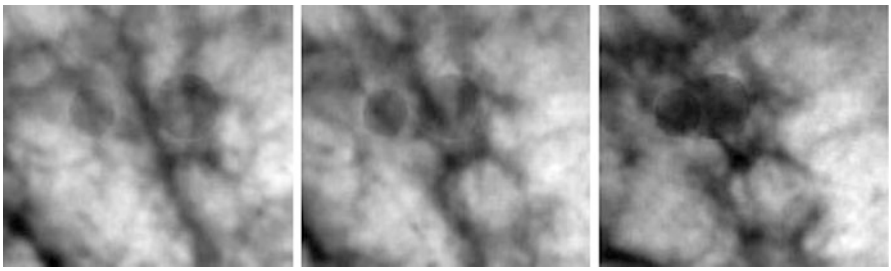


Fig. 12 Extracted regions of sample projection images

the values of the second derivatives. Furthermore, the Newton step at iteration k in Eq. (29) is just the normal equations formulation of the least squares problem

$$\min_{\mathbf{s}_k} \left\| \mathbf{W}_k^{\frac{1}{2}} \mathbf{A}\mathbf{s}_k - \mathbf{W}_k^{-\frac{1}{2}} \mathbf{v}_k \right\|_2 \tag{36}$$

where $\mathbf{W}_k^{\frac{1}{2}} = \text{diag}\left(\mathbf{w}^{\frac{1}{2}}\right)$. For solving the Newton system, CGLS can be used to solve (36) inexactly. Furthermore, regularization for the outer problem is achieved by early termination of the iterative optimization method.

The example illustrated here comes from a true volume of size $128 \times 128 \times 128$ whose values range between 0 and 100, representing the percentage of glandular tissue present in the voxel. Then 21 projection images were taken at equally spaced angles, within an angular range from $-30°$ to $30°$ at $3°$ intervals, using the typical geometry for breast tomosynthesis, illustrated in Fig. 1. Each 2D projection image contains 150×200 pixels. Subimages of three of these projections can be found in Fig. 12.

The original object represented a portion of a patient breast with mean compressed breast thickness of size $6.4 \times 6.4 \times 6.4$ cm, and the detector was 7.5×10 cm. The source to detector distance at $0°$ was set to 66 cm, and the distance from the center of rotation to detector was 0 cm. The incident X-ray spectrum was produced by a rhodium target with a tube voltage of 28 kVp and an added rhodium filter of 25 μm thickness, discretized to consist of 47 different energy levels, from 5.0 to 28 keV, in 0.5 keV steps.

For the reconstruction algorithms, the ray trace matrix $\mathbf{A}_\theta$ for each projection angle was computed using a cone beam model, and an initial guess of the volume was a uniform image with all voxel values set to 50, meaning half glandular and half adipose tissue. The reconstructed volume consisted of $128 \times 128 \times 40$ voxels with a voxel size of $500 \times 500\,\mu\text{m} \times 1.6$ mm. Furthermore, additive Poisson noise was included in the projection images so that there was a relative noise level of approximately 1 %. Some slices of the true volume can be found in Fig. 13.

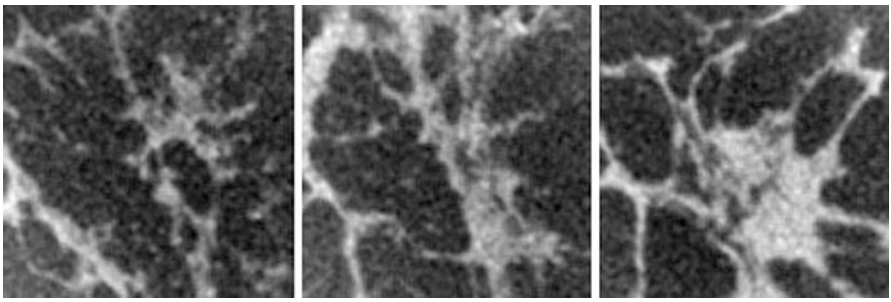


Fig. 13 Sample slices from the original breast volume

Table 1 Convergence results for polyenergetic tomosynthesis reconstruction

Gradient descent method				
Iteration	Rel. objective	Rel. gradient	Rel. error	
0	7.033e − 04	1.0000	0.4034	
1	6.771e − 04	0.8755	0.3562	
5	6.586e − 04	0.2731	0.2948	
10	6.585e − 04	0.0641	0.2762	
25	6.551e − 04	0.0314	0.2386	
50	6.548e − 04	0.0104	0.2237	
Newton-CG				
Iteration	Rel. objective	Rel. gradient	Rel. error	CGLS iterations
0	7.033e − 04	1.0000	0.4034	-
1	6.587e − 04	0.2525	0.2814	5
2	6.550e − 04	0.0398	0.2293	9
3	6.547e − 04	0.0065	0.2075	22
4	6.547e − 04	0.0013	0.2014	50
5	6.547e − 04	0.0009	0.2003	50

Recall that the goal of digital tomosynthesis is to reconstruct an approximation of the 3D volume, $\mathbf{x}$, given the set of projection images $\mathbf{b}_\theta, \theta = 1, 2, \ldots, n_\theta$. Using the above likelihood function, the problem has been reformulated as a nonlinear optimization problem for which standard numerical optimization schemes can be applied. A gradient descent, Newton-CG, and LBFGS algorithm are investigated as methods to solve this problem, and early termination of the iterative method produces a regularized solution.

Results presented in Table 1 include the iteration, the relative objective function value, the relative gradient value, and the relative error for the 3D volume for two iterative algorithms. The relative error can be computed as $\frac{\|\mathbf{x}_k - \mathbf{x}_{\text{exact}}\|_2}{\|\mathbf{x}_{\text{exact}}\|_2}$, where $\mathbf{x}_k$ is the reconstructed volume at the kth iteration. For the inexact Newton-CG algorithm, the stopping criterion used for CGLS on the inner problem (36) was a residual tolerance of 0.17 and a maximum number of 50 iterations. The number of CGLS

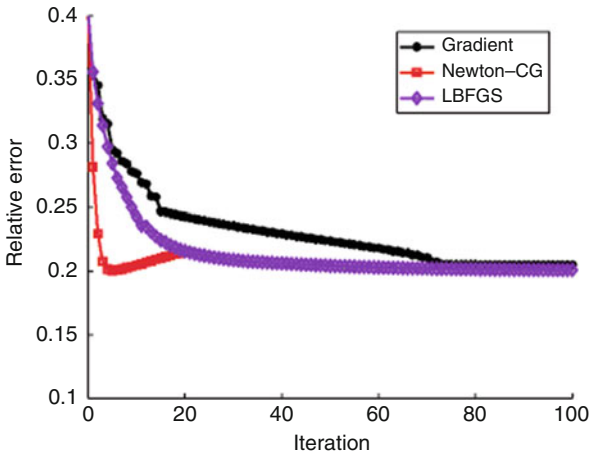


Fig. 14 Plot of relative errors

iterations reported for the inner problem at each Newton-CG iteration can be found in the last column of the table. It is worth mentioning here that many parameters such as the number of inner and outer iterations rely heavily on heuristics that may or may not be provided by the application. In any case, appropriate parameters should be used in order to ensure that nonlinearity and ill-posedness of the problem are addressed.

Since each Newton-CG iteration requires the solution of a linear system, it is difficult to present a fair comparison of reconstruction algorithms. In terms of computational effort, the most computationally burdensome aspect of the reconstruction is the matrix-vector and matrix-transpose-vector multiplications with ray trace matrix, $\mathbf{A}$. Each function and gradient evaluation of the likelihood function requires a total of three "ray trace" multiplications (two for the function evaluation and one more for the gradient), and a multiplication operation with the Hessian (or its transpose) only requires two "ray trace" multiplications. Furthermore, a backtracking line search strategy is used to ensure sufficient descent at each iteration of the optimization scheme. The Cauchy point [77] is used as an initial guess in the line search scheme, thus requiring another multiplication with the Hessian. Thus, the computational cost and timing for, say, one Newton-CG iteration with 50 inner CG iterations with the Hessian is not equivalent to 50 gradient descent iterations.

Another important remark is that although the image errors in Table 1 decrease in the early iterations, these errors eventually increase. This is illustrated in the later Newton-CG iterations in Fig. 14, where plots of the relative errors per iteration for the three algorithms are presented. From Fig. 14, it is evident that the gradient descent method is slow to converge. On the contrary, Newton methods can compute a good approximation very quickly, but corruption from errors occurs quickly as well. Although LBFGS is typically used for problems where the Hessian cannot be computed directly, this approach seems to offer a good balance between fast convergence and slow semi-convergence behavior. An important remark is that

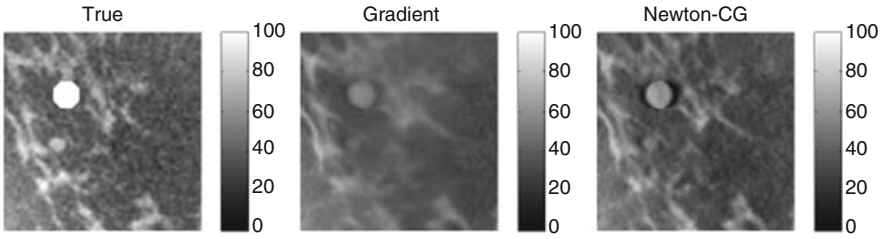


Fig. 15 Comparison of slices from the reconstructed volumes computed after 3 iterations of Newton-CG algorithm and 15 iterations of gradient descent

direct regularization techniques can also be used to regularize the problem, but appropriate regularization operators and good regularization parameter selection methods for this problem are still topics of current research. Thus, regularization via early termination of the iterations is the approach followed here.

For a comparison of images, Fig. 15 contains corresponding slices from the reconstructed volumes after three Newton-CG iterations and 15 gradient descent iterations, each requiring approximately 80 matrix-vector operations. It is evident that the Newton-CG reconstruction has more fine details and more closely resembles the true image slice.

Although nonlinear inverse problems can be difficult to analyze, there are a variety of scientific applications such as polyenergetic digital breast tomosynthesis that require methods for computing approximate solutions. Iterative methods with regularization via early termination can be a good choice, but proper preconditioning techniques may be needed to accelerate the algorithms and good heuristics are required.

5 Conclusion

Large-scale inverse problems arise in many imaging applications. The examples in this chapter illustrate the range of difficulties (from linear to nonlinear) that can be encountered and the issues that must be addressed when designing algorithms. It is important to emphasize that the literature in this field is vast and that this presentation is far from being a complete survey. However, the techniques discussed in this chapter can be used as a foundation on which to learn more about the subject.

The study of inverse problems continues to be an extremely active field of research. Although linear inverse problems have been fairly well studied, some fundamental questions still need to be addressed and many open problems remain. For example, in hybrid algorithms, simple filtering methods (e.g., truncated SVD or standard Tikhonov regularization) and standard regularization parameter choice methods (e.g., discrepancy principle or GCV) are typically used to regularize the projected problem. Some work has been done to generalize this (see, e.g., [59]), but extensions to more sophisticated filtering algorithms and parameter choice methods should be investigated. In addition, the development of novel algorithmic

implementations and software is necessary for running existing algorithms on state-of-the-art computing technologies, as is the development of techniques for uncertainty quantification. Another area of active research for the solution of linear and nonlinear inverse problems is sparse reconstruction schemes, where regularization enforces some structure to be sparse in a certain basis, that is, represented with only a few nonzero coefficients.

As discussed in sections "Separable Inverse Problems" and "Nonlinear Inverse Problems," there are many open problems related to solving nonlinear inverse problems. For example, in the case of the variable projection Gauss-Newton method, a thorough study of its regularization and convergence properties remains to be done, especially in the context of an iteration-dependent regularization parameter. For more general nonlinear problems, issues that need to be addressed include analyzing the sensitivity of the Jacobian and Hessian matrices, as well as determining appropriate merit functions for selecting step lengths. In nonlinear optimization, difficulties arise because convexity of the objective function cannot be guaranteed, so algorithms can become trapped in local minima. More work also needs to be done in the area of regularization parameter choice methods for nonlinear problems and appropriate stopping criteria for iterative methods. For a further discussion of open problems for nonlinear inverse problems, see [28, 29].

Finally, it should be noted that many open problems are given in the context of the application, such as determining appropriate constraints and regularization operators for the problem. Future directions are often motivated by the application, and many of these questions can be found in application-specific references; see, for example, [17]. With such varied and widespread applications, large-scale inverse problems continue to be a thriving research interest in the mathematics, computer science, and image processing communities.

Acknowledgments We would like to thank Eldad Haber, University of British Columbia, and Per Christian Hansen, Technical University of Denmark, for carefully reading the first draft of this chapter. Their comments and suggestions helped to greatly improve our presentation. The research of J. Chung is supported by the US National Science Foundation (NSF) under grant DMS-0902322. The research of J. Nagy is supported by the US National Science Foundation (NSF) under grant DMS-0811031, and by the US Air Force Office of Scientific Research (AFOSR) under grant FA9550-09-1-0487.

Cross-References

▸ Compressive Sensing
▸ Duality and Convex Programming
▸ EM Algorithms
▸ Inverse Scattering
▸ Iterative Solution Methods
▸ Linear Inverse Problems
▸ Numerical Methods and Applications in Total Variation Image Restoration
▸ Optical Imaging

▸ Photoacoustic and Thermoacoustic Tomography: Image Formation Principles
▸ Regularization Methods for Ill-Posed Problems
▸ Statistical Methods in Imaging
▸ Tomography
▸ Total Variation in Imaging

References

1. Andrews, H.C., Hunt, B.R.: Digital Image Restoration. Prentice-Hall, Englewood Cliffs (1977)
2. Bachmayr, M., Burger, M.: Iterative total variation schemes for nonlinear inverse problems. Inverse Prob. **25**, 105004 (2009)
3. Bardsley, J.M.: An efficient computational method for total variation-penalized Poisson likelihood estimation. Inverse Prob. Imaging **2**(2), 167–185 (2008)
4. Bardsley, J.M.: Stopping rules for a nonnegatively constrained iterative method for illposed Poisson imaging problems. BIT **48**(4), 651–664 (2008)
5. Bardsley, J.M., Vogel, C.R.: A nonnegatively constrained convex programming method for image reconstruction. SIAM J. Sci. Comput. **25**(4), 1326–1343 (2003)
6. Barzilai, J., Borwein, J.M.: Two-point step size gradient methods. IMA J. Numer. Anal. **8**(1), 141–148 (1988)
7. Björck, Å.: A bidiagonalization algorithm for solving large and sparse ill-posed systems of linear equations. BIT **28**(3), 659–670 (1988)
8. Björck, Å.: Numerical Methods for Least Squares Problems. SIAM, Philadelphia (1996)
9. Björck, Å., Grimme, E., van Dooren, P.: An implicit shift bidiagonalization algorithm for ill-posed systems. BIT **34**(4), 510–534 (1994)
10. Brakhage, H.: On ill-posed problems and the method of conjugate gradients. In: Engl, H.W., Groetsch, C.W. (eds.) Inverse and Ill-Posed Problems, pp. 165–175. Academic, Boston (1987)
11. Calvetti, D., Reichel, L.: Tikhonov regularization of large linear problems. BIT **43**(2), 263–283 (2003)
12. Calvetti, D., Somersalo, E.: Introduction to Bayesian Scientific Computing. Springer, New York (2007)
13. Candès, E.J., Romberg, J.K., Tao, T.: Stable signal recovery from incomplete and inaccurate measurements. Commun. Pure Appl. Math. **59**(8), 1207–1223 (2006)
14. Carasso, A.S.: Direct blind deconvolution. SIAM J. Appl. Math. **61**(6), 1980–2007 (2001)
15. Chadan, K., Colton, D., Päivärinta, L., Rundell, W.: An Introduction to Inverse Scattering and Inverse Spectral Problems. SIAM, Philadelphia (1997)
16. Chan, T.F., Shen, J.: Image Processing and Analysis: Variational, PDE, Wavelet, and Stochastic Methods. SIAM, Philadelphia (2005)
17. Cheney, M., Borden, B.: Fundamentals of Radar Imaging. SIAM, Philadelphia (2009)
18. Chung, J., Haber, E., Nagy, J.: Numerical methods for coupled super-resolution. Inverse Prob. **22**(4), 1261–1272 (2006)
19. Chung, J., Nagy, J.: An efficient iterative approach for large-scale separable nonlinear inverse problems. SIAM J. Sci. Comput. **31**(6), 4654–4674 (2010)
20. Chung, J., Nagy, J., Sechopoulos, I.: Numerical algorithms for polyenergetic digital breast tomosynthesis reconstruction. SIAM J. Imaging Sci. **3**(1), 133–152 (2010)
21. Chung, J., Nagy, J.G., O'Leary, D.P.: A weighted GCV method for Lanczos hybrid regularization. Elec. Trans. Numer. Anal. **28**, 149–167 (2008)
22. Chung, J., Sternberg, P., Yang, C.: High performance 3-d image reconstruction for molecular structure determination. Int. J. High Perform. Comput. Appl. **24**(2), 117–135 (2010)
23. De Man, B., Nuyts, J., Dupont, P., Marchal, G., Suetens, P.: An iterative maximumlikelihood polychromatic algorithm for CT. IEEE Trans. Med. Imaging **20**(10), 999–1008 (2001)

24. Diaspro, A., Corosu, M., Ramoino, P., Robello, M.: Two-photon excitation imaging based on a compact scanning head. IEEE Eng. Med. Biol. **18**(5), 18–30 (1999)
25. Dobbins, J.T., III, Godfrey, D.J.: Digital X-ray tomosynthesis: current state of the art and clinical potential. Phys. Med. Biol. **48**(19), R65–R106 (2003)
26. Easley, G.R., Healy, D.M., Berenstein, C.A.: Image deconvolution using a general ridgelet and curvelet domain. SIAM J. Imaging Sci. **2**(1), 253–283 (2009)
27. Elad, M., Feuer, A.: Restoration of a single superresolution image from several blurred, noisy, and undersampled measured images. IEEE Trans. Image Process. **6**(12), 1646–1658 (1997)
28. Engl, H.W., Hanke, M., Neubauer, A.: Regularization of Inverse Problems. Kluwer, Dordrecht (2000)
29. Engl, H.W., Kügler, P.: Nonlinear inverse problems: theoretical aspects and some industrial applications. In: Capasso, V., Périaux, J. (eds.) Multidisciplinary Methods for Analysis Optimization and Control of Complex Systems, pp. 3–48. Springer, Berlin (2005)
30. Engl, H.W., Kunisch, K., Neubauer, A.: Convergence rates for Tikhonov regularisation of nonlinear ill-posed problems. Inverse Prob. **5**(4), 523–540 (1989)
31. Engl, H.W., Louis, A.K., Rundell, W. (eds.): Inverse Problems in Geophysical Applications. SIAM, Philadelphia (1996)
32. Eriksson, J., Wedin, P.: Truncated Gauss-Newton algorithms for ill-conditioned nonlinear least squares problems. Optim. Meth. Softw. **19**(6), 721–737 (2004)
33. Faber, T.L., Raghunath, N., Tudorascu, D., Votaw, J.R.: Motion correction of PET brain images through deconvolution: I. Theoretical development and analysis in software simulations. Phys. Med. Biol. **54**(3), 797–811 (2009)
34. Figueiredo, M.A.T, Nowak, R.D., Wright, S.J.: Gradient projection for sparse reconstruction: application to compressed sensing and other inverse problems. IEEE J. Sel. Top. Signal Process. **1**(4), 586–597 (2007)
35. Frank, J.: Three-Dimensional Electron Microscopy of Macromolecular Assemblies. Oxford University Press, New York (2006)
36. Golub, G.H., Heath, M., Wahba, G.: Generalized cross-validation as a method for choosing a good ridge parameter. Technometrics **21**(2), 215–223 (1979)
37. Golub, G.H., Luk, F.T., Overton, M.L.: A block Lanczos method for computing the singular values and corresponding singular vectors of a matrix. ACM Trans. Math. Softw. **7**(2), 149–169 (1981)
38. Golub, G.H., Pereyra, V.: The differentiation of pseudo-inverses and nonlinear least squares problems whose variables separate. SIAM J. Numer. Anal. **10**(2), 413–432 (1973)
39. Golub, G.H., Pereyra, V.: Separable nonlinear least squares: the variable projection method and its applications. Inverse Prob. **19**, R1–R26 (2003)
40. Haber, E., Ascher, U.M., Oldenburg, D.: On optimization techniques for solving nonlinear inverse problems. Inverse Prob. **16**(5), 1263–1280 (2000)
41. Haber, E., Oldenburg, D.: A GCV based method for nonlinear ill-posed problems. Comput. Geosci. **4**(1), 41–63 (2000)
42. Hammerstein, G.R., Miller, D.W., White, D.R., Masterson, M.E., Woodard, H.Q., Laughlin, J.S.: Absorbed radiation dose in mammography. Radiology **130**(2), 485–491 (1979)
43. Hanke, M.: Conjugate gradient type methods for ill-posed problems. Pitman research notes in mathematics, Longman Scientific & Technical, Harlow (1995)
44. Hanke, M.: Limitations of the L-curve method in ill-posed problems. BIT **36**(2), 287–301 (1996)
45. Hanke, M.: On Lanczos based methods for the regularization of discrete ill-posed problems. BIT **41**(5), 1008–1018 (2001)
46. Hansen, P.C.: Analysis of discrete ill-posed problems by means of the L-curve. SIAM Rev. **34**(4), 561–580 (1992)
47. Hansen, P.C.: Numerical tools for analysis and solution of Fredholm integral equations of the first kind. Inverse Prob. **8**(6), 849–872 (1992)
48. Hansen, P.C.: Regularization tools: a MATLAB package for analysis and solution of discrete ill-posed problems. Numer. Algorithms **6**(1), 1–35 (1994)

49. Hansen, P.C.: Rank-Deficient and Discrete Ill-Posed Problems. SIAM, Philadelphia (1998)
50. Hansen, P.C.: Discrete Inverse Problems: Insight and Algorithms. SIAM, Philadelphia (2010)
51. Hansen, P.C., Nagy, J.G., O'Leary, D.P.: Deblurring Images: Matrices, Spectra and Filtering. SIAM, Philadelphia (2006)
52. Hansen, P.C., O'Leary, D.P.: The use of the L-curve in the regularization of discrete ill-posed problems. SIAM J Sci Comput 14(6):1487–1503 (1993)
53. Hardy, J.W.: Adaptive optics. Sci. Am. **270**(6), 60–65 (1994)
54. Hofmann, B.: Regularization of nonlinear problems and the degree of ill-posedness. In: Anger, G., Gorenflo, R., Jochmann, H., Moritz, H., Webers, W. (eds.) Inverse Problems: Principles and Applications in Geophysics, Technology, and Medicine. Akademie Verlag, Berlin (1993)
55. Hohn, M., Tang, G., Goodyear, G., Baldwin, P.R., Huang, Z., Penczek, P.A., Yang, C., Glaeser, R.M., Adams, P.D., Ludtke, S.J.: SPARX, a new environment for Cryo-EM image processing. J. Struct. Biol. **157**(1), 47–55 (2007)
56. Jain, A.K.: Fundamentals of Digital Image Processing. Prentice-Hall, Englewood Cliffs (1989)
57. Kang, M.G., Chaudhuri, S.: Super-resolution image reconstruction. IEEE Signal Process. Mag. **20**(3), 19–20 (2003)
58. Kaufman, L.: A variable projection method for solving separable nonlinear least squares problems. BIT **15**(1), 49–57 (1975)
59. Kilmer, M.E., Hansen, P.C., Español, M.I.: A projection-based approach to general-form Tikhonov regularization. SIAM J. Sci. Comput. **29**(1), 315–330 (2007)
60. Kilmer, M.E., O'Leary, D.P.: Choosing regularization parameters in iterative methods for ill-posed problems. SIAM J. Matrix. Anal. Appl. **22**(4), 1204–1221 (2001)
61. Landweber, L.: An iteration formula for Fredholm integral equations of the first kind. Am. J. Math. **73**(3), 615–624 (1951)
62. Larsen, R.M.: Lanczos bidiagonalization with partial reorthogonalization. PhD thesis, Department of Computer Science, University of Aarhus, Denmark (1998)
63. Lawson, C.L., Hanson, R.J.: Solving Least Squares Problems. SIAM, Philadelphia (1995)
64. Löfdahl, M.G.: Multi-frame blind deconvolution with linear equality constraints. In: Bones, P.J., Fiddy, M.A., Millane, R.P. (eds.) Image Reconstruction from Incomplete Data II, Seattle, vol. 4792–21, pp. 146–155. SPIE (2002)
65. Lohmann, A.W., Paris, D.P.: Space-variant image formation. J. Opt. Soc. Am. **55**(8), 1007–1013 (1965)
66. Marabini, R., Herman, G.T., Carazo, J.M.: 3D reconstruction in electron microscopy using ART with smooth spherically symmetric volume elements (blobs). Ultramicroscopy **72**(1–2), 53–65 (1998)
67. Matson, C.L., Borelli, K., Jefferies, S., Beckner, C.C., Jr., Hege, E.K., Lloyd-Hart, M.: Fast and optimal multiframe blind deconvolution algorithm for high-resolution groundbased imaging of space objects. Appl. Opt. **48**(1), A75–A92 (2009)
68. McNown, S.R., Hunt, B.R.: Approximate shift-invariance by warping shift-variant systems. In: Hanisch, R.J., White, R.L. (eds.) The Restoration of HST Images and Spectra II, pp. 181–187. Space Telescope Science Institute, Baltimore (1994)
69. Miller, K.: Least squares methods for ill-posed problems with a prescribed bound. SIAM J. Math. Anal. **1**(1), 52–74 (1970)
70. Modersitzki, J.: Numerical Methods for Image Registration. Oxford University Press, Oxford (2004)
71. Morozov, V.A.: On the solution of functional equations by the method of regularization. Sov. Math. Dokl. **7**, 414–417 (1966)
72. Nagy, J.G., O'Leary, D.P.: Fast iterative image restoration with a spatially varying PSF. In: Luk, F.T. (ed.) Advanced Signal Processing: Algorithms, Architectures, and Implementations VII, San Diego, vol. 3162, pp. 388–399. SPIE (1997)
73. Nagy, J.G., O'Leary, D.P.: Restoring images degraded by spatially-variant blur. SIAM J. Sci. Comput. **19**(4), 1063–1082 (1998)

74. Natterer, F.: The Mathematics of Computerized Tomography. SIAM, Philadelphia (2001)
75. Natterer, F., Wübbeling, F.: Mathematical Methods in Image Reconstruction. SIAM, Philadelphia (2001)
76. Nguyen, N., Milanfar, P., Golub, G.: Efficient generalized cross-validation with applications to parametric image restoration and resolution enhancement. IEEE Trans. Image Process. **10**(9), 1299–1308 (2001)
77. Nocedal, J., Wright, S.: Numerical Optimization. Springer, New York (1999)
78. O'Leary, D.P., Simmons, J.A.: A bidiagonali-zation-regularization procedure for large scale discretizations of ill-posed problems. SIAM J. Sci. Stat. Comput. **2**(4), 474–489 (1981)
79. Osborne, M.R.: Separable least squares, variable projection, and the Gauss-Newton algorithm. Elec. Trans. Numer. Anal. **28**, 1–15 (2007)
80. Paige, C.C., Saunders, M.A.: Algorithm 583 LSQR: sparse linear equations and least squares problems. ACM Trans. Math. Softw. **8**(2), 195–209 (1982)
81. Paige, C.C., Saunders, M.A.: LSQR: an algorithm for sparse linear equations and sparse least squares. ACM Trans. Math. Softw. **8**(1), 43–71 (1982)
82. Penczek, P.A., Radermacher, M., Frank, J.: Three-dimensional reconstruction of single particles embedded in ice. Ultramicroscopy **40**(1), 33–53 (1992)
83. Phillips, D.L.: A technique for the numerical solution of certain integral equations of the first kind. J. Assoc. Comput. Mach. **9**(1), 84–97 (1962)
84. Raghunath, N., Faber, T.L., Suryanarayanan, S., Votaw, J.R.: Motion correction of PET brain images through deconvolution: II. Practical implementation and algorithm optimization. Phys. Med. Biol. **54**(3), 813–829 (2009)
85. Rudin, L.I., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. Physica D **60**, 259–268 (1992)
86. Ruhe, A., Wedin, P.: Algorithms for separable nonlinear least squares problems. SIAM Rev. **22**(3), 318–337 (1980)
87. Saad, Y.: On the rates of convergence of the Lanczos and the block-Lanczos methods. SIAM J. Numer. Anal. **17**(5), 687–706 (1980)
88. Saban, S.D., Silvestry, M., Nemerow, G.R., Stewart, P.L.: Visualization of α-helices in a 6-Ångstrom resolution cryoelectron microscopy structure of adenovirus allows refinement of capsid protein assignments. J. Virol. **80**(24), 49–59 (2006)
89. Tikhonov, A.N.: Regularization of incorrectly posed problems. Sov. Math. Dokl. **4**, 1624–1627 (1963)
90. Tikhonov, A.N.: Solution of incorrectly formulated problems and the regularization method. Sov. Math. Dokl. **4**, 1035–1038 (1963)
91. Tikhonov, A.N., Arsenin, V.Y.: Solutions of Ill-Posed Problems. Winston, Washington (1977)
92. Tikhonov, A.N., Leonov, A.S., Yagola, A.G.: Nonlinear Ill-Posed Problems, vol. 1–2. Chapman and Hall, London (1998)
93. Trussell, H.J., Fogel, S.: Identification and restoration of spatially variant motion blurs in sequential images. IEEE Trans. Image Process. **1**(1), 123–126 (1992)
94. Tsaig, Y., Donoho, D.L.: Extensions of compressed sensing. Signal Process. **86**(3), 549–571 (2006)
95. Varah, J.M.: Pitfalls in the numerical solution of linear ill-posed problems. SIAM J. Sci. Stat. Comput. **4**(2), 164–176 (1983)
96. Vogel, C.R.: Optimal choice of a truncation level for the truncated SVD solution of linear first kind integral equations when data are noisy. SIAM J. Numer. Anal. **23**(1), 109–117 (1986)
97. Vogel, C.R.: An overview of numerical methods for nonlinear ill-posed problems. In: Engl, H.W., Groetsch, C.W. (eds.) Inverse and Ill-Posed Problems, pp. 231–245. Academic, Boston (1987)
98. Vogel, C.R.: Non-convergence of the L-curve regularization parameter selection method. Inverse Prob. **12**(4), 535–547 (1996)
99. Vogel, C.R.: Computational Methods for Inverse Problems. SIAM, Philadelphia (2002)
100. Wagner, F.C., Macovski, A., Nishimura, D.G.: A characterization of the scatter pointspread-function in terms of air gaps. IEEE Trans. Med. Imaging **7**(4), 337–344 (1988)

Regularization Methods for Ill-Posed Problems

Jin Cheng and Bernd Hofmann

Contents

1 Introduction 92
2 Theory of Direct Regularization Methods 93
Tikhonov Regularization in Hilbert Spaces with Quadratic Misfit and Penalty Terms 96
Variational Regularization in Banach Spaces with Convex Penalty Term 98
Some Specific Results for Hilbert Space Situations 104
Further Convergence Rates Under Variational Inequalities 106
3 Examples 108
4 Conclusion 118
Cross-References 118
References 118

Abstract

In this chapter are outlined some aspects of the mathematical theory for direct regularization methods aimed at the stable approximate solution of nonlinear ill-posed inverse problems. The focus is on Tikhonov type variational regularization applied to nonlinear ill-posed operator equations formulated in Hilbert and Banach spaces. The chapter begins with the consideration of the classical approach in the Hilbert space setting with quadratic misfit and penalty terms, followed by extensions of the theory to Banach spaces and present assertions on convergence and rates concerning the variational regularization with general convex penalty terms. Recent results refer to the interplay between solution smoothness and nonlinearity conditions expressed by variational inequalities.

J. Cheng (✉)
School of Mathematical Sciences, Fudan University, Shanghai, China
e-mail: jcheng@fudan.edu.cn

B. Hofmann
Faculty of Mathematics, Technische Universität Chemnitz, Chemnitz, Germany
e-mail: bernd.hofmann@mathematik.tu-chemnitz.de; hofmannb@mathematik.tu-chemnitz.de

© Springer Science+Business Media New York 2015

O. Scherzer (ed.), *Handbook of Mathematical Methods in Imaging*,
DOI 10.1007/978-1-4939-0790-8_3

Six examples of parameter identification problems in integral and differential equations are given in order to show how to apply the theory of this chapter to specific inverse and ill-posed problems.

1 Introduction

This chapter will be devoted to direct regularization methods – theory and examples – for the solution of inverse problems formulated as nonlinear ill-posed operator equations

$$F(x) = y\,, \tag{1}$$

where the forward operator $F : \mathcal{D}(F) \subseteq X \to Y$ with domain $\mathcal{D}(F)$ maps between infinite dimensional normed linear spaces X and Y, which are Banach spaces or Hilbert spaces, with norms $\|\cdot\|$. The symbol $\langle \cdot, \cdot \rangle$ designates inner products in Hilbert spaces and dual pairings in Banach spaces. Moreover, the symbols $\to$ and $\rightharpoonup$ denote the norm convergence and weak convergence, respectively, in such spaces. It is well known that the majority of inverse problems are ill-posed in the sense of Hadamard, i.e., at least one of the following difficulties occurs:

(i) Equation (1) has no solution in $\mathcal{D}(F)$ if the exact right-hand side y is replaced by a perturbed element y^δ (noisy data) satisfying the inequality

$$\left\| y^\delta - y \right\| \le \delta \tag{2}$$

with noise level $\delta > 0$.

(ii) The solution to Eq. (1) is not uniquely determined in $\mathcal{D}(F)$.

(iii) The solution to Eq. (1) is unstable with respect to perturbations, i.e., for $x^\delta \in \mathcal{D}(F)$ with $F\left(x^\delta\right) = y^\delta$ and (2) the norm deviation $\left\| x^\delta - x \right\|$ may be arbitrarily large. In other words, the possibly multivalued inverse operator F^{-1} fails to be continuous.

Since for nonlinear equations the local behavior of solutions is of main interest, the aspect of local ill-posedness according to [64] is focused on numerous considerations. An operator equation (1) is called locally ill-posed at some solution point $x^\dagger \in \mathcal{D}(F)$ if for any ball $B_\rho\left(x^\dagger\right)$ with center $x^\dagger$ and an arbitrarily small radius $\rho > 0$ there exist infinite sequences $\{x_n\} \subset B_\rho\left(x^\dagger\right) \cap \mathcal{D}(F)$ such that

$$F(x_n) \to F\left(x^\dagger\right) \quad \text{in } Y, \qquad \text{but} \quad x_n \not\to x^\dagger \quad \text{in } X \quad \text{as } n \to \infty\,.$$

In case of local ill-posedness, $x^\dagger$ cannot be identified arbitrarily precise by noisy data y^δ even if the noise level δ is arbitrarily small. The aspect of local ill-posedness involves both the non-injectivity of F around $x^\dagger$ corresponding with (ii) and the local instability of (1) corresponding with (iii) in Hadamard's sense. Wide classes of inverse problems that have smoothing, for example, compact, forward operators F lead to locally ill-posed situations.

To overcome the ill-posedness and local ill-posedness of Eq. (1), in particular to compensate the instability of solutions with respect to small changes in the right-hand side expressing a deficit of information in the data with respect to the solution to be determined, regularization methods have to be used for the stable approximate solution of Eq. (1) whenever only noisy data are given. The basic idea of regularization is to replace the ill-posed original problem by a well-posed and stable neighboring problem. A regularization parameter $\alpha > 0$ controls the trade-off between closeness of the neighboring problem expressed by small values α and high stability of the auxiliary problem expressed by large values α. In the former case the approximate solutions are too unstable, whereas in the latter case, approximate solutions are too far from the original one. On the other hand, the loss of information in the data caused by smoothing properties of the forward operator F can be diminished when external a priori information is exploited. This can be done by the choice of appropriate structure in the neighboring problems.

If the forward operator F and hence the operator equation (1) is linear, then in Hilbert spaces a comprehensive and rather complete regularization theory, including a general regularization schema and a well-established collection of methods, assertions on stability, convergence, and convergence rates, is available since more than 20 years; see [10, 35, 77–79, 84]. For recent progress of regularization theory applied to linear ill-posed problems, please refer to the papers [16, 22, 31, 57, 61, 80, 85, 86, 90]. It is well known that inverse problems aimed at the identification of parameter functions in differential equations or boundary conditions from observations of state variables are in general nonlinear even if the differential equations are linear. The nonlinearity of F, however, makes the construction and the use of regularization methods more complicated and diversified; please refer to [9, 26, 27, 33, 38, 45, 49, 56, 66, 71, 81, 87] for more details. Furthermore, there have been recent significant progress in regularization theory for ill-posed problems formulated in Banach spaces (see, e.g., [22, 60, 96, 97, 102, 103, 107, 109]). In this chapter, the focus is on direct regularization methods for the stable approximate solution of nonlinear ill-posed operator equations formulated in Hilbert and Banach spaces, where regularized solutions mostly are solutions of variational problems. The functional to be minimized over a set of admissible solutions contains a regularization parameter $\alpha > 0$ which has to be controlled in an appropriate manner. An alternative way of regularization is the solution of (1) for noisy data y^δ by an iteration process, where the stopping criterion, frequently depending on δ, plays the role of the regularization parameter. For iterative solution methods, please refer to the corresponding chapter of this book and to the monograph [8, 76].

2 Theory of Direct Regularization Methods

In contrast to the classical treatment of linear ill-posed problems, where regularized solutions $x_\alpha^\delta = R_\alpha y^\delta$, i.e., stable approximate solutions to Eq. (1) under the noise model (2), are obtained by applying bounded linear operators $R_\alpha : Y \to X$ to the data y^δ for regularization parameters $\alpha > 0$, such explicit approach fails if either

in (1) (a) the forward operator F is nonlinear, (b) the domain $\mathcal{D}(F)$ is not a linear subspace of X, or (c) the mapping $y^\delta \mapsto x_\alpha^\delta$ is continuous but nonlinear for all $\alpha > 0$ even if F is linear. All three sources of nonlinearity make it necessary to define the regularized solutions in an implicit manner. The preferred approach of direct regularization methods is variational regularization, where regularized solutions x_α^δ are minimizers of the functional

$$\Phi(x) := \mathcal{S}\left(F(x), y^\delta\right) + \alpha \mathcal{R}(x) \tag{3}$$

by assuming that $\mathcal{S}$ is a nonnegative misfit functional measuring the discrepancy between $F(x)$ and the data y^δ, moreover $\alpha > 0$ is the regularization parameter, and $\mathcal{R}$ with domain $\mathcal{D}(\mathcal{R}) := \{x \in X : \mathcal{R}(x) < \infty\}$ is a nonnegative stabilizing functional with small values for elements x being reliable and large values for improbable x. Since the origins of this method go back to the work of A. N. Tikhonov and his collaborators (see [114, 115]), this method is often called Tikhonov type regularization. For example, please refer to the monographs [7, 8, 10, 35, 54, 59, 68, 73, 92, 107, 109, 116, 117] and to the papers [82, 83, 98, 118], which contribute to the theory and practice of this kind of regularization.

Besides the standard version

$$\mathcal{S}\left(F(x), y^\delta\right) = \left\| F(x) - y^\delta \right\|^p, \qquad p \geq 1, \tag{4}$$

which is mostly used in combination with the noise model (2) in Banach spaces, specific noise models like Poisson noise or a stochastic background suggest alternative choices for the misfit term $\mathcal{S}$ like Kullback-Leibler or other divergences. With respect to imaging, e.g., deblurring, image reconstruction, image registration, and partial differential equations occurring there, different chapters of the monographs [2, 15, 39, 41, 91, 94, 106, 107] and the papers [12, 23, 40, 93, 120] motivate and discuss regularized solutions x_α^δ as well as different choices of functionals S and $\mathcal{R}$. On the other hand, the minimizers of (3) play also an important role in the treatment of statistical inverse problems by Bayesian methods, maximum a posteriori estimation (MAP), and penalized maximum likelihood estimation (see, e.g., [74]), where in some cases the penalty term $\mathcal{R}(x)$ can even be determined by a priori information when the solution x is a realization of a randomized state variable.

A typical property of ill-posed equations is that minimizing $\mathcal{S}\left(F(x), y^\delta\right)$ alone is very sensitive to data changes and yields in most cases highly oscillating minimizers. For example, the least-squares approach $\left\| F(x) - y^\delta \right\|^2 \to \min$ as preferred method in Hilbert spaces shows such behavior. Therefore, the regularization parameter $\alpha > 0$ in the variational problem $\Phi(x) \to \min$ controls the trade-off between optimal data fitting with unstable solutions if α is near zero and a high level of stability and sympathy but larger misfit for the approximate solution if α is more far from zero. The set of admissible solutions in the process of minimizing Φ from (3) is the intersection

$$\mathcal{D} := \mathcal{D}(F) \cap \mathcal{D}(\mathcal{R})$$

of the domains of F and $\mathcal{R}$. For obtaining a regularized solution x_α^δ to a nonlinear inverse problem, a nonlinear and frequently non-convex optimization problem has to be solved, since either the functional Φ or the set $\mathcal{D}(F)$ can be non-convex. As a consequence, for the numerical treatment of direct regularization methods in combination with discretization approaches, iterative procedures are also required. In this context, the details have been omitted, and the reader is referred to the monographs [25, 35, 107, 116] and to the sample [16, 31, 69, 72, 101] of papers from a comprehensive set of publications on numerical approaches.

The appropriate choice of α is one of the most serious tasks in regularization, where a priori choices $\alpha = \alpha(\delta)$ and a posteriori choices $\alpha = \alpha\left(\delta, y^\delta\right)$ have to be distinguished. For a priori choices the decay rate of $\alpha(\delta) \to 0$ as $\delta \to 0$ is prescribed with the goal that regularized solutions converge to a solution of (1), i.e., $x_{\alpha(\delta)}^\delta \to x^\dagger$ as $\delta \to 0$. Such convergence can be arbitrarily slow depending on smoothness properties of $x^\dagger$. To obtain convergence rates $\left\| x_{\alpha(\delta)}^\delta - x^\dagger \right\| = \mathcal{O}(\varphi(\delta))$ as $\delta \to 0$, that means a uniform convergence for some nonnegative increasing rate function $\varphi(\delta)$ with $\varphi(0) = 0$, additional conditions on $x^\dagger$, so-called source conditions, have to be satisfied. In contrast to a priori choices, an a posteriori choice of the regularization parameter α takes into account the present data y^δ and tries to equilibrate the noise level δ and the deviation between $F\left(x_{\alpha(\delta, y^\delta)}^\delta\right)$ and y^δ. By the discrepancy principle, as the most prominent approach, α is chosen originally such that $\left\| F\left(x_{\alpha(\delta, y^\delta)}\right)^\delta - x^\dagger \right\| = C\delta$ with some constant $C \geq 1$ whenever (4) and (2) are supposed. Various discussions, generalizations, and improvements of the discrepancy principle can be found in [3, 5, 69, 92, 113]. If δ is not known sufficiently well, then heuristic methods for choosing $\alpha = \alpha\left(y^\delta\right)$ can be exploited as the quasioptimality principle, the L-curve method, and others (see, e.g., [7, 35, 55]). They have theoretical drawbacks, since convergence fails in worst case situations, but the utility of those methods for many classes of applications is beyond controversy.

The choices of $\mathcal{S}$, $\mathcal{R}$, and $\alpha = \alpha\left(\delta, y^\delta\right)$ should be realized such that the following questions Q1–Q4 can be answered in a positive manner:

- Q1: Do minimizers x_α^δ of the functional Φ from (3) exist for all $\alpha > 0$ and $y^\delta \in Y$?
- Q2: Do the minimizers x_α^δ for fixed $\alpha > 0$ stably depend on the data y^δ?
- Q3: Is there a convergence $x_{\alpha(\delta, y^\delta)}^\delta \to x^\dagger$ to a solution $x^\dagger$ of (1) if under (2) $\delta \to 0$?
- Q4: Are there sufficient conditions imposed on $x^\dagger$ for obtaining convergence rates $\left\| x_\alpha^\delta - x^\dagger \right\| = \mathcal{O}(\varphi(\delta))$ as $\delta \to 0$? In this context, $\varphi : (0, \infty) \to (0, \infty)$ is an index function, which means that φ is continuous and strictly increasing with $\lim_{\delta \to +0} \varphi(\delta) = 0$.

Sometimes the requirement of norm convergence is too strong. If, for instance, the penalty functional attains the form $\mathcal{R}(x) = \|x\|^q$, $q > 0$, then $\mathcal{R}$ is stabilizing only in the sense of a topology, which is weaker than the norm topology in X. Precisely, the level sets $\{x \in X : \mathcal{R}(x) \leq c\}$ are weakly sequentially compact in X

if X is a Hilbert space or a reflexive Banach space. In such case, weak convergence $x^{\delta}_{\alpha(\delta,y^{\delta})} \rightharpoonup x^{\dagger}$ of regularized solutions is a reasonable requirement. This leads to norm convergence of regularized solutions if the Radon-Riesz property is satisfied; see [107, 109]. On the other hand, it may be useful to replace the norm as a measure for the error of regularization by alternative measures $E\left(x^{\delta}_{\alpha}, x^{\dagger}\right)$, preferably the Bregman distance if $\mathcal{R}$ is a convex functional and X is a Banach space; see [22, 41, 50, 58, 60].

Tikhonov Regularization in Hilbert Spaces with Quadratic Misfit and Penalty Terms

In Hilbert spaces X and Y, quadratic Tikhonov regularization with the functional

$$\Phi(x) := \left\| F(x) - y^{\delta} \right\|^2 + \alpha \left\| x - x^* \right\|^2 \tag{5}$$

to be minimized over $\mathcal{D} = \mathcal{D}(F)$ is the most prominent variant of variational regularization of nonlinear ill-posed operator equations, for which the complete theory with respect to questions Q1–Q4 was elaborated 20 years ago (see [37, 110]). For a comprehensive presentation including the convergence rates results, please refer to [35, Chapter 10].

For some initial guess or reference element, $x^* \in X$ minimizers of (5) tend to approximate x^*-minimum norm solutions $x^{\dagger}$ to (1) for which

$$\left\| x^{\dagger} - x^* \right\| = \min \left\{ \|x - x^*\| : \ F(x) = y, \ x \in \mathcal{D}(F) \right\}.$$

Note that x^*-minimum norm solutions need not exist. In case of existence they need not be uniquely determined. However, under the following assumption, the existence of a solution $x^{\dagger} \in \mathcal{D}(F)$ to (1) implies the existence of an x^*-minimum norm solution (see [107, Lemma 3.2]).

Assumption 1. *1. The operator $F : \mathcal{D}(F) \subseteq X \to Y$ maps between Hilbert spaces X and Y with a nonempty domain $\mathcal{D}(F)$.*
2. F is weakly sequentially closed, i.e., weak convergence of the sequences $x_n \rightharpoonup x_0$ and $F(x_n) \rightharpoonup y_0$ with $x_n \in \mathcal{D}(F)$, $x_0 \in X$, $y_0 \in Y$ implies $x_0 \in \mathcal{D}(F)$ and $F(x_0) = y_0$.

For checking item 2 of Assumption 1, it is important to know that in the case of weakly closed and convex domains $\mathcal{D}(F)$, the weak continuity of F, i.e., $x_n \rightharpoonup x_0$ implies $F(x_n) \rightharpoonup F(x_0)$, is a sufficient condition. Moreover, the following proposition (see [35, Section 10.2]) answers the questions Q1–Q3 in a positive manner.

Proposition 1. *Under Assumption 1 the functional (5) has a minimizer $x^{\delta}_{\alpha} \in \mathcal{D}(F)$ for all $\alpha > 0$ and $y^{\delta} \in Y$. For fixed $\alpha > 0$ and a sequence $y_n \to y^{\delta}$, every infinite sequence $\{x_n\}$ of minimizers to the associated functionals*

$$\Phi_n(x) := \|F(x) - y_n\|^2 + \alpha \|x - x^*\|^2 \tag{6}$$

has a convergent subsequence, and all limits of such subsequences are minimizers x_α^δ *of (5). Whenever the a priori parameter choice* $\alpha = \alpha(\delta) > 0$ *for* $\delta > 0$ *satisfies*

$$\alpha(\delta) \to 0 \quad \textit{and} \quad \frac{\delta^2}{\alpha(\delta)} \to 0\,, \quad \textit{as} \ \ \delta \to 0\,,$$

and if (1) has a solution in $\mathcal{D}(F)$, $\delta_n \to 0$ *is a sequence of noise levels with corresponding data* $y_n = y^{\delta_n}$ *such that* $\|y_n - y\| \le \delta_n$, *then every associated sequence* $\{x_n\}$ *of minimizers to (6) has a convergent subsequence, and all limit elements are* x^**-minimum norm solutions* $x^\dagger$ *of (1).*

An answer to question Q4 concerning convergence rates is given by the following theorem along the lines of [35, Theorem 10.4].

Theorem 2. *In addition to Assumption 1, let* $\mathcal{D}(F)$ *be convex and* $x^\dagger$ *be an* x^**-minimum norm solution to (1) such that*

$$\left\|F(x) - F\left(x^\dagger\right) - A\left(x - x^\dagger\right)\right\| \le \frac{L}{2}\left\|x - x^\dagger\right\|^2 \quad \textit{for all} \ \ x \in \mathcal{D}(F) \cap B_\rho\left(x^\dagger\right) \tag{7}$$

holds for a ball $B_\rho\left(x^\dagger\right)$ *with sufficiently large radius* $\rho > 0$, *a constant* $L > 0$, *and a bounded linear operator* $A : X \to Y$ *satisfying a source condition*

$$x^\dagger - x^* = A^* w\,, \tag{8}$$

where $A^* : Y \to X$ *is the adjoint operator to* A *and* $w \in Y$ *is some source element fulfilling the smallness condition*

$$L\,\|w\| < 1\,. \tag{9}$$

Then the error estimate

$$\left\|x_\alpha^\delta - x^\dagger\right\| \le \frac{\delta + \alpha\|w\|}{\sqrt{\alpha}\sqrt{1 - L\|w\|}}$$

and, for the a priori parameter choice $\underline{c}\delta \le \alpha(\delta) \le \overline{c}\delta$ *with constants* $0 < \underline{c} \le \overline{c} < \infty$, *the convergence rate*

$$\left\|x_{\alpha(\delta)}^\delta - x^\dagger\right\| = \mathcal{O}\left(\sqrt{\delta}\right) \quad \textit{as} \ \ \delta \to 0 \tag{10}$$

are obtained.

The operator A in (7) must be considered as a linearization of F at the point $x^\dagger$ in the sense of a Gâteaux or Fréchet derivative $F'\left(x^\dagger\right)$. The condition (7) characterizes

the structure of nonlinearity of the forward operator F in a neighborhood of $x^\dagger$. If the Fréchet derivative $F'(x)$ exists and is locally Lipschitz continuous around $x^\dagger$ with Lipschitz constant $L > 0$, then (7) is fulfilled.

For further convergence rate results of Tikhonov regularization in Hilbert spaces with quadratic penalty term, please refer, for example, to [64,75,88,89,95,105,111–113].

Variational Regularization in Banach Spaces with Convex Penalty Term

In Banach spaces X and Y, the wide variety of variational regularization realized by minimizing the functional Φ from (3) allows establishing a priori information about the noise model and the solution $x^\dagger$ to be determined in a more sophisticated manner than Tikhonov regularization in Hilbert spaces with quadratic misfit and penalty terms. Knowledge of the specific situation motivates the selection of the functionals $\mathcal{S}$ and $\mathcal{R}$, where norm powers (4) are considered here as misfit functional, which, for example, simplifies the numerical treatment of minimization problems if Y is a Lebesgue space $Y = L^p(\Omega)$ or a Sobolev space $Y = W^{l,p}(\Omega)$ with $1 \le p \le \infty$, $\Omega \subset \mathbb{R}^k$. Please refer to the papers [6, 32, 44, 51, 53, 63, 99] for a further discussion of alternative misfit functionals $\mathcal{S}$. In most cases convex penalty functionals $\mathcal{R}$ are preferred. An important class of penalty functionals form the norm powers $\mathcal{R}(x) := \|x\|_{\tilde{X}}^q$, $q > 0$, $x \in \tilde{X}$, where as an alternative to the standard case $X = \tilde{X}$ the space $\tilde{X}$ can also be chosen as a dense subspace of X with stronger norm, e.g., $X = L^q(\Omega)$, $\tilde{X} = W^{l,q}(\Omega)$, $l = 1, 2, \ldots$. To reconstruct non-smooth solutions $x^\dagger$, the exponent q can be chosen smaller than two, for example, close to one or $q = 1$ if the solution is assumed to be sparse. To recover solutions for which the expected smoothness is low, also penalty terms $\mathcal{R}(x) = TV(x)$ are frequently applied, partly in a modified manner, where $TV(x) = \int_\Omega |\nabla x|$ expresses the total variation of the function x (see, e.g., [1, 13, 107, 118, 119]). For specific applications in imaging (see [107, Chapter 5]) and to handle sparsity of solutions (see [3,18,19,50,52,122] and [107, Section 3.3]), the systematic use of non-convex misfit and penalty functionals can be appropriate, in particular $0 < q < 1$ for the norm power penalties $\mathcal{R}$. In the sequel, however, the focus in this section in combination with the noise model (2) is upon the functional

$$\Phi(x) := \left\| F(x) - y^\delta \right\|^p + \alpha \mathcal{R}(x), \qquad 1 \le p < \infty, \tag{11}$$

with a convex penalty functional $\mathcal{R}$ to be minimized over $\mathcal{D} = \mathcal{D}(F) \cap \mathcal{D}(\mathcal{R})$ yielding minimizers x_α^δ.

Assumption 3. *1. The operator $F : \mathcal{D}(F) \subseteq X \to Y$ maps between reflexive Banach spaces X and Y with duals X^* and Y^*, respectively.*

2. *F is weakly sequentially closed, $\mathcal{D}(F)$ is convex and weakly closed, and the intersection $\mathcal{D} = \mathcal{D}(F) \cap \mathcal{D}(\mathcal{R})$ is nonempty.*
3. *The functional $\mathcal{R}$ is convex and weakly sequentially lower semicontinuous.*
4. *The functional $\mathcal{R}$ is stabilizing, which means that for every $\alpha > 0,\ c \geq 0$, and for the exact right-hand side y of (1), the level sets*

$$\mathcal{M}_\alpha(c) := \{x \in \mathcal{D} : \|F(x) - y\|^p + \alpha \mathcal{R}(x) \leq c\} \tag{12}$$

are weakly sequentially pre-compact in the following sense: every infinite sequence $\{x_n\}$ in $\mathcal{M}_\alpha(c)$ has a subsequence, which is weakly convergent in X to some element from X.
5. *The operator F possesses for all $x_0 \in \mathcal{D}$ a one-sided directional derivative with Gâteaux derivative-like properties, i.e., for every $x_0 \in \mathcal{D}$ there is a bounded linear operator $F'(x_0) : X \to Y$ such that*

$$\lim_{t \to +0} \frac{1}{t} \left(F(x_0 + t(x - x_0)) - F(x_0)\right) = F'(x_0)(x - x_0)$$

is valid for all $x \in \mathcal{D}(F)$.

Under Assumption 3 existence and stability of regularized solutions x_α^δ can be shown (see [60, §3]), i.e., the questions Q1–Q2 above get a positive answer. In Banach spaces regularization errors are estimated by upper bounds of $E\left(x_\alpha^\delta, x^\dagger\right)$, where E denotes an appropriate nonnegative error measure. Besides the norm deviation

$$E\left(x_\alpha^\delta, x^\dagger\right) = \left\|x_\alpha^\delta - x^\dagger\right\|, \tag{13}$$

which is the standard error measure in Hilbert spaces, in Banach spaces and for convex functionals $\mathcal{R}$ with subdifferential $\partial\mathcal{R}$, the Bregman distance

$$E\left(x_\alpha^\delta, x^\dagger\right) = D_\xi\left(x_\alpha^\delta, x^\dagger\right) := \mathcal{R}\left(x_\alpha^\delta\right) - \mathcal{R}\left(x^\dagger\right) - \left\langle \xi, x_\alpha^\delta - x^\dagger \right\rangle \tag{14}$$

at $x^\dagger \in \mathcal{D}_B(\mathcal{R}) \subseteq X$ and $\xi \in \partial\mathcal{R}\left(x^\dagger\right) \subseteq X^*$ is frequently used as error measure, where the set $\mathcal{D}_B(\mathcal{R}) := \{x \in \mathcal{D}(\mathcal{R}) : \partial\mathcal{R}(x) \neq \emptyset\}$ represents the Bregman domain. An element $x^\dagger \in \mathcal{D}$ is called an $\mathcal{R}$-minimizing solution to (1) if

$$\mathcal{R}\left(x^\dagger\right) = \min\{\mathcal{R}(x) : F(x) = y,\ x \in \mathcal{D}\} < \infty\,.$$

Such $\mathcal{R}$-minimizing solutions exist under Assumption 3 if (1) has a solution $x \in \mathcal{D}$. For given $\alpha_{\max} > 0$ let $x^\dagger$ denote an $\mathcal{R}$-minimizing solution of (1). By setting

$$\rho := 2^{p-1}\alpha_{\max}\left(1 + \mathcal{R}\left(x^\dagger\right)\right), \tag{15}$$

it holds $x^\dagger \in \mathcal{M}_{\alpha_{\max}}(\rho)$ and there exists some $\delta_{\max} > 0$ such that

$$x^{\delta}_{\alpha(\delta)} \in \mathcal{M}_{\alpha_{\max}}(\rho) \qquad \text{for all} \qquad 0 \le \delta \le \delta_{\max}\,.$$

Along the lines of [107, Section 3.2], [109, Chapters 3 and 4], and [3, 17, 58, 60, 62, 65], there are presented in the following some results on the regularization theory for that setting.

Under an a priori parameter choice $\alpha = \alpha(\delta) > 0$ satisfying

$$\alpha(\delta) \to 0 \quad \text{and} \quad \frac{\delta^p}{\alpha(\delta)} \to 0\,, \quad \text{as } \delta \to 0\,,$$

a positive answer to question Q3 can be given and weak convergence $x^{\delta}_{\alpha(\delta)} \rightharpoonup x^{\dagger}$ as $\delta \to 0$ (for subsequences in analogy to Proposition 1) of regularized solutions to $\mathcal{R}$-minimizing solutions $x^{\dagger}$ is shown. For stronger convergence results concerning the norm convergence, please refer, for example, to Proposition 3.32 in [107].

Taking into account the advantages of a posteriori choices of the regularization parameter $\alpha > 0$, it is of interest to select such parameter choice rules $\alpha = \alpha\left(\delta, y^{\delta}\right)$ which obey the conditions

$$\alpha\left(\delta, y^{\delta}\right) \to 0 \qquad \text{and} \qquad \frac{\delta^p}{\alpha\left(\delta, y^{\delta}\right)} \to 0 \qquad \text{as} \qquad \delta \to 0. \tag{16}$$

Such study was performed for the sequential discrepancy principle which will be introduced in the following. The basis of this variety of discrepancy principle is, for prescribed $0 < q < 1$ and $\alpha_0 > 0$, a sequence

$$\Delta_q := \{\alpha_j > 0 : \ \alpha_j = q^j\,\alpha_0, \quad j \in \mathbb{Z}\} \tag{17}$$

of regularization parameters and the specification of some constant $\tau > 1$.

Definition 1 (sequential discrepancy principle). For fixed $\delta > 0$ and $y^{\delta} \in Y$, it is said that $\alpha = \alpha\left(\delta, y^{\delta}\right) \in \Delta_q$ is chosen according to the sequential discrepancy principle if

$$\left\| F\left(x^{\delta}_{\alpha}\right) - y^{\delta} \right\| \le \tau\,\delta < \left\| F\left(x^{\delta}_{\alpha/q}\right) - y^{\delta} \right\|. \tag{18}$$

Due to Assumption 3, the set $X_{\min} := \{x \in \mathcal{D} : \mathcal{R}(x) = \mathcal{R}_{\min}\}$ is nonempty for the value $\mathcal{R}_{\min} := \inf\limits_{x \in \mathcal{D}} \mathcal{R}(x) \ge 0$ and so is $Y_{\min} := F(X_{\min})$, where

$$F(S) := \{\hat{y} \in Y : \ \hat{y} = F(x),\ x \in S\}$$

is defined for subsets $S \subseteq \mathcal{D}(F)$. Moreover, there is an element $x_{\min} \in X_{\min}$ such that $\operatorname{dist}\left(y^{\delta}, Y_{\min}\right) := \inf\limits_{x \in X_{\min}} \left\| F(x) - y^{\delta} \right\| = \left\| F(x_{\min}) - y^{\delta} \right\|$. Using this notation, the Proposition 2 below still needs the following two definitions concerning the exact penalization veto and requirements on data compatibility, respectively.

Definition 2 (exact penalization veto). It is said that the exact penalization veto is satisfied for $y \in F(\mathcal{D})$ if, for all $\alpha > 0$, any minimizer $x^\dagger$ of the functional

$$\Phi_0(x) := \|F(x) - y\|^p + \alpha \mathcal{R}(x)$$

over $\mathcal{D}$, which is simultaneously an $\mathcal{R}$-minimizing solution of (1), belongs to $X_{\min}$.

Definition 3 (compatible data). For $y \in F(\mathcal{D})$ and prescribed $\tau > 1$, it is said that there is data compatibility if there is some $\delta_{\max} > 0$ such that for all data $y^\delta \in Y$ fulfilling (2), the condition

$$\tau\delta < \operatorname{dist}\left(y^\delta, Y_{\min}\right) \qquad \text{for all} \qquad 0 < \delta \leq \delta_{\max}$$

is satisfied.

In the paper [3], there have been formulated sufficient conditions for fulfilling the exact penalization veto, which is mostly the case if $p > 1$ for the norm exponent in the misfit term of the Tikhonov functionals Φ and Φ_0. Sufficient conditions for obtaining data compatibility can also be found in that paper. The following proposition is a direct consequence of the corresponding studies in [3] for an α-selection according to the sequential discrepancy principle.

Proposition 2. *Let for the exact right-hand side $y \in F(\mathcal{D})$ of Eq. (1) the exact penalization veto be satisfied and assume for prescribed $\tau > 1$ data compatibility. Then there is some $\delta_{\max} > 0$ such that $\alpha = \alpha\left(\delta, y^\delta\right)$ can be chosen according to the sequential discrepancy principle for all $0 < \delta \leq \delta_{\max}$. Moreover, this parameter choice satisfies condition (16), and consequently weak convergence $x^\delta_{\alpha(\delta,y^\delta)} \rightharpoonup x^\dagger$ as $\delta \to 0$ occurs in the sense of subsequences to $\mathcal{R}$-minimizing solutions $x^\dagger$ of (1) with the limit condition $\lim_{\delta\to 0} \mathcal{R}\left(x^\delta_{\alpha(\delta,y^\delta)}\right) = \mathcal{R}\left(x^\dagger\right)$.*

The weak convergence of regularized solutions $x^\delta_{\alpha(\delta,y^\delta)} \rightharpoonup x^\dagger$ in the sense of subsequences from Proposition 2 is not worth too much. Even if the Radon-Riesz property, which ensures that $x^\delta_{\alpha(\delta,y^\delta)} \rightharpoonup x^\dagger$ and $\mathcal{R}\left(x^\delta_{\alpha(\delta,y^\delta)}\right) \to \mathcal{R}\left(x^\dagger\right)$ lead to $x^\delta_{\alpha(\delta,y^\delta)} \to x^\dagger$ as $\delta \to 0$, allows amplifying this to norm convergence, this convergence can be arbitrarily slow for awkward solutions $x^\dagger$. To get more, namely, a uniform error estimate for classes of solutions $x^\dagger$ in the sense of convergence rates with an index function φ determining the rate (see question Q4 above), additional requirements must be imposed on all elements $x^\dagger$ from the class under consideration. In practice, one can restrict the considerations to rate functions φ which are concave. The abovementioned requirements are always smoothness conditions. Precisely, the $\mathcal{R}$-minimizing solutions $x^\dagger$ to be approximated by regularized solutions

must have a certain level of smoothness with respect to F. For nonlinear forward operators F also the specific structure of nonlinearity influences the rate function φ.

With respect to operator properties, first the concept of a degree of nonlinearity from [58] is exploited:

Definition 4 (degree of nonlinearity in Banach space). Let $c_1 \geq 0$, $c_2 \geq 0$, and $c_1 + c_2 > 0$. Then F is said to be nonlinear of degree (c_1, c_2) for the Bregman distance $D_\xi\left(\cdot, x^\dagger\right)$ of $\mathcal{R}$ at a solution $x^\dagger \in \mathcal{D}_B(\mathcal{R}) \subseteq X$ of (1) with $\xi \in \partial\mathcal{R}\left(x^\dagger\right) \subseteq X^*$ if there is a constant $K > 0$ such that

$$\left\| F(x) - F\left(x^\dagger\right) - F'\left(x^\dagger\right)\left(x - x^\dagger\right) \right\| \leq K \left\| F(x) - F\left(x^\dagger\right) \right\|^{c_1} D_\xi\left(x, x^\dagger\right)^{c_2} \tag{19}$$

for all $x \in \mathcal{M}_{\alpha_{\max}}(\rho)$.

On the other hand, the solution smoothness of $x^\dagger$ in combination with a well-defined degree of nonlinearity can be expressed in an efficient manner by variational inequalities

$$\left\langle \xi, x^\dagger - x \right\rangle \leq \beta_1 D_\xi\left(x, x^\dagger\right) + \beta_2 \left\| F(x) - F\left(x^\dagger\right) \right\|^\kappa \qquad \text{for all} \quad x \in \mathcal{M}_{\alpha_{\max}}(\rho) \tag{20}$$

with some $\xi \in \partial\mathcal{R}(x^\dagger)$, two multipliers $0 \leq \beta_1 < 1$, $\beta_2 \geq 0$, and an exponent $\kappa > 0$ for obtaining convergence rates. The subsequent theorem (for a proof see [65]) shows the utility of such variational inequalities for ensuring convergence rates in variational regularization (for more details in the case $\kappa = 1$, see also [60] and [107, Section 3.2]).

Theorem 4. *For regularized solutions x_α^δ minimizing Φ from (11) with $p > 1$ under Assumption 3 and provided that there is an $\mathcal{R}$-minimizing solution $x^\dagger \in \mathcal{D}_B(\mathcal{R})$, the convergence rate*

$$E\left(x_{\alpha(\delta)}^\delta, x^\dagger\right) = \mathcal{O}\left(\delta^\kappa\right) \quad \text{as} \quad \delta \to 0 \tag{21}$$

is valid for the Bregman distance (14) as error measure and for an a priori parameter choice $\alpha(\delta) \asymp \delta^{p-\kappa}$ if there exist an element $\xi \in \partial\mathcal{R}\left(x^\dagger\right)$ and constants $0 \leq \beta_1 < 1$, $\beta_2 \geq 0$ such that the variational inequality (20) holds for some $0 < \kappa \leq 1$ and with ρ from (15).

This result which is based on Young's inequality can immediately be extended to the situation $0 < \kappa < p \leq 1$. Moreover, the situation $\kappa = p \leq 1$ characterizes the exact penalization case. For noisy data and $\kappa = p \leq 1$, it holds $D_\xi\left(x_{\alpha_0}^\delta, x^\dagger\right) = \mathcal{O}\left(\delta^p\right)$ as $\delta \to 0$ for a regularization parameter $\alpha = \alpha_0$ which is fixed but sufficiently small (see [22]). An extension of such results to convergence rates of higher order is outlined in [97].

To verify different situations for the exponent $\kappa > 0$, the setting is restricted as follows:

Assumption 5. *In addition to Assumption 3, assume for an $\mathcal{R}$-minimizing solution $x^\dagger$ to (1):*

1. *The operator F is Gâteaux differentiable in $x^\dagger$ with the Gâteaux derivative $F'\left(x^\dagger\right)$.*
2. *The functional $\mathcal{R}$ is Gâteaux differentiable in $x^\dagger$ with the Gâteaux derivative $\xi = \mathcal{R}'\left(x^\dagger\right) \in X^*$; hence, the subdifferential $\partial\mathcal{R}\left(x^\dagger\right) = \{\xi\}$ is a singleton.*

The following proposition (see [65, Proposition 4.3]) shows that exponents $\kappa > 1$ in the variational inequality (20) under Assumption 5 in principle cannot occur.

Proposition 3. *Under the Assumption 5 the variational inequality (20) cannot hold with $\xi = \mathcal{R}'\left(x^\dagger\right) \neq 0$ and multipliers $\beta_1, \beta_2 \geq 0$ whenever $\kappa > 1$.*

Now the following proposition will highlight the borderline case $\kappa = 1$ and the cross connections between variational inequalities and source conditions for the Banach space setting. Moreover, in the next subsection, the interplay with (8) and generalized source condition can be discussed.

Proposition 4. *Under Assumption 5 the following two assertions hold:*

(a) The validity of a variational inequality

$$\left\langle \xi, x^\dagger - x \right\rangle \leq \beta_1 D_\xi\left(x, x^\dagger\right) + \beta_2 \left\| F(x) - F\left(x^\dagger\right) \right\| \quad \text{for all} \quad x \in \mathcal{M}_{\alpha_{\max}}(\rho) \tag{22}$$

for $\xi = \mathcal{R}'\left(x^\dagger\right)$ and two multipliers $\beta_1, \beta_2 \geq 0$ implies the source condition

$$\xi = F'\left(x^\dagger\right)^* w, \qquad w \in Y^*. \tag{23}$$

(b) Let F be nonlinear of degree $(0, 1)$ for the Bregman distance $D_\xi(\cdot, x^\dagger)$ of $\mathcal{R}$ at $x^\dagger$, i.e.,

$$\left\| F(x) - F\left(x^\dagger\right) - F'\left(x^\dagger\right)\left(x - x^\dagger\right) \right\| \leq K\, D_\xi\left(x, x^\dagger\right) \tag{24}$$

holds for a constant $K > 0$ and all $x \in \mathcal{M}_{\alpha_{\max}}(\rho)$. Then the source condition (23) together with the smallness condition

$$K \|w\|_{Y^*} < 1 \tag{25}$$

implies the validity of a variational inequality (22) with $\xi = \mathcal{R}'\left(x^\dagger\right)$ and multipliers $0 \leq \beta_1 = K\|w\|_{Y^*} < 1$, $\beta_2 = \|w\|_{Y^*} \geq 0$.

Sufficient conditions for the validity of a variational inequality (20) with fractional exponents $0 < \kappa < 1$ are formulated in [58] based on the method of approximate source conditions using appropriate distance functions that measure the degree of violation of the source condition (23) for the solution $x^\dagger$. Assertions on convergence rates for that case can be made when the degree of nonlinearity is such that $c_1 > 0$ as the next proposition shows.

Proposition 5. *Under Assumption 5 let F be nonlinear of degree (c_1, c_2) with $0 < c_1 \le 1$, $0 \le c_2 < 1$, $c_1 + c_2 \le 1$ for the Bregman distance $D_\xi\left(\cdot, x^\dagger\right)$ of $\mathcal{R}$ at $x^\dagger$, i.e.,*

$$\left\| F(x) - F\left(x^\dagger\right) - F'\left(x^\dagger\right)\left(x - x^\dagger\right)\right\| \le K \left\| F(x) - F\left(x^\dagger\right)\right\|^{c_1} D_\xi\left(u, x^\dagger\right)^{c_2} \tag{26}$$

holds for a constant $K > 0$ and all $x \in \mathcal{M}_{\alpha_{\max}}(\rho)$. Then the source condition (23) immediately implies the validity of a variational inequality (20) with

$$\kappa = \frac{c_1}{1 - c_2}, \tag{27}$$

$\xi = \mathcal{R}'\left(x^\dagger\right)$, and multipliers $0 \le \beta_1 < 1$, $\beta_2 \ge 0$.

Some Specific Results for Hilbert Space Situations

The abstract concepts of the last subsection for the Tikhonov regularization with quadratic functionals under Assumption 5, but with Assumption 1 in Hilbert spaces, will be illustrated next. For

$$\mathcal{R}(x) := \|x - x^*\|^2$$

the $\mathcal{R}$-minimizing solutions and the classical x^*-minimum norm solutions coincide. Moreover, it holds $\mathcal{D} = \mathcal{D}(F)$ and for ξ and $D_\xi(\tilde{x}, x)$ the simple structure

$$\xi = 2\left(x^\dagger - x^*\right) \qquad \text{and} \qquad D_\xi(\tilde{x}, x) = \|\tilde{x} - x\|^2$$

with Bregman domain $\mathcal{D}_B(\mathcal{R}) = X$. Then the source condition (23) attains the form (8) with $A = F'\left(x^\dagger\right)/2$.

To focus on the distinguished character of the setting for Hilbert spaces X and Y, the Definition 4 will be specified as follows:

Definition 5 (Degree of nonlinearity in Hilbert space). Let $c_1 \ge 0$, $c_2 \ge 0$, and $c_1 + c_2 > 0$. Define F to be nonlinear of degree (c_1, c_2) at a solution $x^\dagger \in \mathcal{D}(F)$ of (1) if there is a constant $K > 0$ such that

$$\left\| F(x) - F\left(x^\dagger\right) - F'\left(x^\dagger\right)\left(x - x^\dagger\right)\right\| \;\le\; K\;\left\|F(x) - F\left(x^\dagger\right)\right\|^{c_1}\;\left\|x - x^\dagger\right\|^{2c_2} \tag{28}$$

for all $x \in \mathcal{M}_{\alpha_{\max}}(\rho)$.

Furthermore, in Hilbert spaces Hölder source conditions

$$\xi = \left(F'\left(x^\dagger\right)^* F'(x^\dagger)\right)^{\eta/2} v, \qquad v \in X, \tag{29}$$

can be formulated that allow expressing a lower level of solution smoothness of $x^\dagger$ for $0 < \eta < 1$ compared to the case $\eta = 1$, where (29) is equivalent to

$$\xi = F'\left(x^\dagger\right)^* w, \qquad w \in Y$$

(cf. condition (8) in Theorem 2). For that situation of lower smoothness, the following theorem (see [65, Proposition 6.6]) complements the Theorem 2.

Theorem 6. *Under the Assumption 5 let the operator F mapping between the Hilbert spaces X and Y be nonlinear of degree (c_1, c_2) at $x^\dagger$ with $c_1 > 0$ and let with $\mathcal{R}(x) = \|x - x^*\|^2$ the element $\xi = 2\left(x^\dagger - x^*\right)$ satisfy the Hölder source condition (29). Then the variational inequality (20) holds with exponent*

$$\kappa \;=\; \min\left\{\frac{2\eta c_1}{1+\eta(1-2c_2)}\,,\;\frac{2\eta}{1+\eta}\right\}\,, \qquad 0 < \eta \le 1, \tag{30}$$

for all $x \in \mathcal{M}_{\alpha_{\max}}(\rho)$ and multipliers $0 \le \beta_1 < 1$, $\beta_2 \ge 0$. Consequently, for regularized solutions $x^\delta_{\alpha(\delta)}$ minimizing the functional Φ from (5), the convergence rate

$$E\left(x^\delta_{\alpha(\delta)}, x^\dagger\right) \;=\; \mathcal{O}\left(\delta^{\kappa/2}\right) \quad \text{as} \quad \delta \to 0 \tag{31}$$

holds for the norm distance (13) as error measure and for an a priori parameter choice
$\alpha(\delta) \asymp \delta^{p-\kappa}$.

For parameter identification problems in partial differential equations (cf., e.g., [9, 68]), which can be written as nonlinear operator equations (1) with implicitly given forward operators F, it is difficult to estimate the Taylor remainder $\left\|F(x) - F\left(x^\dagger\right) - F'\left(x^\dagger\right)\left(x - x^\dagger\right)\right\|$ and the variational inequality approach may fail. However, if in addition to the Hilbert space X a densely defined subspace $\tilde{X}$ with stronger norm is considered, then in many applications for all $R > 0$, there hold conditional stability estimates of the form

$$\|x_1 - x_2\| \le K \|F(x_1) - F(x_2)\|^\kappa \quad \text{if} \quad x_i \in \mathcal{D}(F) \cap \tilde{X}, \ \|x_i\|_{\tilde{X}} \le R \ \ (i = 1, 2) \tag{32}$$

with some $0 < \kappa \le 1$ and a constant $K = K(R) > 0$, which may depend on the radius R.

Then along the lines of the paper, [31] the following theorem can be formulated.

Theorem 7. *Let X and Y be Hilbert spaces and let $B : \mathcal{D}(B) \subset X \to X$ be an unbounded injective, positive definite, self-adjoint linear operator with domain $\tilde{X} = \mathcal{D}(B)$ dense in X. Furthermore, let $\tilde{C} > 0$ be a constant such that $\|x\|_{\tilde{X}} := \|Bx\| \ge \tilde{C} \|x\|$ holds for $x \in \tilde{X}$ and $\tilde{X}$ becomes a Hilbert space with norm $\|\cdot\|_{\tilde{X}}$ stronger than the norm in X. For the nonlinear operator $F : \mathcal{D}(F) \subseteq X \to Y$, consider regularized solutions x_α^δ as minimizers over $\mathcal{D} := \mathcal{D}(F) \cap \tilde{X}$ of the functional*

$$\Phi(x) := \left\| F(x) - y^\delta \right\|^2 + \alpha \|x\|_{\tilde{X}}^2 . \tag{33}$$

Moreover, for all $R > 0$ let hold a conditional stability estimate of the form (32) with some $0 < \kappa \le 1$ and a constant $K = K(R) > 0$. Then for a solution $x^\dagger \in \mathcal{D}$ of Eq. (1), the convergence rate

$$E\left(x_{\alpha(\delta)}^\delta, x^\dagger\right) = \mathcal{O}\left(\delta^\kappa\right) \quad \text{as} \quad \delta \to 0 \tag{34}$$

is obtained with the norm distance (13) as error measure and for an a priori parameter choice $\underline{c}\delta^2 \le \alpha(\delta) \le \overline{c}\delta^2$ for constants $0 < \underline{c} \le \overline{c} < \infty$.

The papers [28, 65] show that the convergence rate result of Theorem 7 can be extended to Banach space situations.

Further Convergence Rates Under Variational Inequalities

Returning to the Banach space setting, in Theorem 4, convergence rates were drawn from the variational inequality (20) benefit. For the error measure E from (14) and taking into account that the penalty functional $\mathcal{R}$ is convex, this inequality can be reformulated as

$$\beta\, E\left(x, x^\dagger\right) \le \mathcal{R}(x) - \mathcal{R}\left(x^\dagger\right) + C \left\| F(x) - F\left(x^\dagger\right) \right\| \quad \text{for all} \quad x \in \mathcal{M} ,$$

for constants $0 < \beta \le 1$ and $C > 0$, where $\mathcal{M}$ denotes an appropriate set which contains $x^\dagger$ and all regularized solutions x_α^δ for sufficiently small $\delta > 0$. Such variational inequalities (also called variational range conditions or variational smoothness assumptions) are in the case of general nonnegative error measures E a powerful tool for obtaining convergence rates in Tikhonov type regularization, and please refer to the overview in [42]. In the extended form

$$\beta\, E\left(x, x^{\dagger}\right) \leq \mathcal{R}(x) - \mathcal{R}\left(x^{\dagger}\right) + C\, \varphi\left(\left\|F(x) - F\left(x^{\dagger}\right)\right\|\right) \quad \text{for all} \quad x \in \mathcal{M}\,, \tag{35}$$

with some concave index function φ, such variational inequalities were intensively studied in [62]. Inequalities of this type combine assertions on the solution smoothness of $x^{\dagger}$ with assertions on the structure of nonlinearity of F around $x^{\dagger}$. In [17] it was shown that inequalities of the form (35) are a consequence of nonlinearity conditions

$$\left\|F'\left(x^{\dagger}\right)\left(x - x^{\dagger}\right)\right\| \leq \sigma\left(\left\|F(x) - F\left(x^{\dagger}\right)\right\|\right) \quad \text{for all} \quad x \in \mathcal{M}\,, \tag{36}$$

with concave index functions σ, where φ depends on σ, and both functions even coincide if $x^{\dagger}$ satisfies the source condition

$$\xi = F'\left(x^{\dagger}\right)^{*} w, \qquad w \in Y^{*}\,, \tag{37}$$

for some subgradient $\xi \in \partial\mathcal{R}\left(x^{\dagger}\right)$.

Theorem 8. *Suppose that for $y \in F(\mathcal{D})$ the $\mathcal{R}$-minimizing solution $x^{\dagger} \in \mathcal{D}$ of Eq. (1) obeys for the nonnegative error measure E the variational inequality (35) with some set $\mathcal{M} \subseteq X$, constants $0 < \beta \leq 1$, $C > 0$, and a concave index function φ. Then, for the a priori parameter choice*

$$\alpha(\delta) = \alpha_0 \frac{\delta^p}{\varphi(\delta)}$$

and provided that the Tikhonov-regularized solutions satisfy the condition $x^{\delta}_{\alpha(\delta)} \in \mathcal{M}$ for $0 < \delta \leq \delta_{\max}$, the convergence rate

$$E\left(x^{\delta}_{\alpha(\delta)}, x^{\dagger}\right) = \mathcal{O}\left(\varphi(\delta)\right) \qquad \text{as} \qquad \delta \to 0 \tag{38}$$

holds for arbitrarily chosen $\alpha_0 > 0$ whenever the norm exponent p in (11) is taken from the interval $1 < p < \infty$ and for $0 < \alpha_0 < 1$ if $p = 1$.

The same convergence rate result can also be derived from the variational inequality (35) if the regularization parameter α is chosen according to the sequential discrepancy principle. For linear injective operators F and an ℓ^1-penalty term $\mathcal{R}$, this was exploited in [21] (see also [4]) for obtaining convergence rates under conjectured sparsity constraints when the sparsity fails, but $x^{\dagger}$ is in ℓ^1. Then the decay rate of solution components $x^{\dagger}_k \to 0$ as $k \to \infty$ influences the function φ in (35).

3 Examples

In this section, several examples of parameter identification problems in integral and differential equations will be presented in order to show how to apply the regularization theory outlined above to specific ill-posed and inverse problems. The examples refer either to nonlinear inverse problems, which can be formulated as operator equations (1) with forward operator F mapping from a Hilbert space X to a Hilbert space Y, or to linearizations of such problems, which then appear as linear operator equations. All discussed examples originally represent ill-posed problems in the sense that small data changes may lead to arbitrarily large errors in the solution. If the forward operator F is linear, then this phenomenon can be characterized by the fact that the range of the operator F is a nonclosed subspace in Y. For nonlinear F such simple characterization fails, but a local version of ill-posedness (see [64]) takes place in general. In order to make clear the cross connections to the theory, as in the previous sections, the unknown parameter functions are denoted by x, in particular the exact solution to (1) by $x^\dagger$, and the exact and noisy data are denoted by y and y^δ, respectively. For conciseness, this section restricts to six examples. More examples can be found in the corresponding references of this book.

Example I ((Identification of coefficients in wave equations)). Let $\Omega \subset \mathbb{R}^n$, $n = 1, 2, 3$, be a bounded domain with C^2-boundary $\partial\Omega$. Consider

$$\begin{cases} \frac{\partial^2 u}{\partial t^2}(\zeta, t) = \Delta u(\zeta, t) + x(\zeta) u(\zeta, t), \quad \zeta \in \Omega, \, 0 < t < T, \\ u(\zeta, 0) = a(\zeta), \quad \frac{\partial u}{\partial t}(\zeta, 0) = b(\zeta), \quad \zeta \in \Omega, \\ \frac{\partial u}{\partial \nu}(\zeta, t) = 0, \quad \zeta \in \partial\Omega, \quad 0 < t < T. \end{cases} \tag{39}$$

Here and henceforth $\frac{\partial}{\partial \nu}$ denotes the normal derivative. Fix initial values a and b such that

$$a \in H^3(\Omega), \quad b \in H^2(\Omega), \quad \frac{\partial a}{\partial \nu}\Big|_{\partial\Omega\times(0,T)} = \frac{\partial b}{\partial \nu}\Big|_{\partial\Omega\times(0,T)} = 0. \tag{40}$$

Then for any function $x \in W^{1,\infty}(\Omega)$, there exists a unique solution

$$u(x) = u(x)(\zeta, t) \in C\left([0, T]; H^3(\Omega)\right) \cap C^1\left([0, T]; H^2(\Omega)\right) \cap C^2\left([0, T]; H^1(\Omega)\right)$$

to (39) (see, e.g., [67]).

The inverse problem here consists in the identification of the parameter function $x = x(\zeta)$, $\zeta \in \Omega$, occurring in the hyperbolic partial differential equation based on noisy observations y^δ of time derivatives y of the state variable $[u(x)](\zeta, t)$ on the boundary $(\zeta, t) \in \partial\Omega \times (0, T)$. In addition to (40), assume that

$$T > \min_{\zeta' \in \overline{\Omega}} \max_{\zeta \in \overline{\Omega}} |\zeta - \zeta'| \tag{41}$$

and

$$|a(\zeta)| > 0, \qquad \zeta \in \overline{\Omega}. \tag{42}$$

Moreover, set

$$\mathcal{U}_M = \left\{x \in W^{1,\infty}(\Omega) : \ \|x\|_{W^{1,\infty}(\mathbb{R})} \leq M\right\} \tag{43}$$

for $M > 0$.

In [67], it is proved that there exists a constant $C = C(\Omega, T, a, b, M) > 0$ such that

$$\|x_1 - x_2\|_{L^2(\Omega)} \leq C \left\| \frac{\partial}{\partial t}(u(x_1) - u(x_2)) \right\|_{H^1(\partial\Omega\times(0,T))} \tag{44}$$

for all $x_1, x_2 \in \mathcal{U}_M$.

The forward operator F is defined as a mapping from the space $X = L^2(\Omega)$ to the space $Y = H^1(\partial\Omega \times (0, T))$ according to

$$[F(x)](\zeta, t) := \frac{\partial u(x)}{\partial t}\Big|_{\partial\Omega\times(0,T)}, \quad (\zeta, t) \in \partial\Omega \times (0, T).$$

This is a nonlinear operator mapping between the Hilbert spaces X and Y, and Ill-posedness of the corresponding operator equation can be indicated. However, the estimate (44) shows that this inverse problem possesses good stability properties if the set of admissible solutions is restricted suitably or if the regularization term is chosen in an appropriate manner.

A Tikhonov regularization approach as outlined in the previous sections is useful. If the functional Φ is chosen as

$$\Phi(x) = \left\|F(x) - y^\delta\right\|_Y^2 + \alpha \left\|x - x^*\right\|_X^2,$$

the theory applies. Alternatively, the penalty term can also be chosen by using a stronger norm. In this case, the functional Φ is chosen as

$$\Phi(x) = \left\|F(x) - y^\delta\right\|_Y^2 + \alpha\|x\|_{\tilde{X}}^2.$$

where $\tilde{X} = W^{1,\infty}(\Omega)$.

Then by the conditional stability estimation (44), the convergence rate

$$\left\|x_\alpha^\delta - x^\dagger\right\|_{L^2(\Omega)} = \mathcal{O}(\delta) \quad \text{as} \quad \delta \to 0$$

is obtained with the choice $\alpha = \delta^2$.

Example II (Determination of shapes of boundaries). Using polar coordinates (r, θ), the shape of a boundary in $\mathbb{R}^2$ is identified. For $M > 0$ and $0 < m_0 <$

$m_1 < 1$, set

$$\mathcal{U}_{m_1,M} = \Big\{ x = x(\theta) \in C^2[0,2\pi] : \ \frac{d^k x}{d\theta^k}(0) = \frac{d^k x}{d\theta^k}(2\pi),\ k = 0,1,2,$$
$$\|x\|_{C^2[0,2\pi]} \le M, \quad \|x\|_{C[0,2\pi]} \le m_1 \Big\}$$

and

$$Q_{m_0} = \big\{ x \in C^2[0,2\pi] : \ x(\theta) \ge m_0,\ 0 \le \theta \le 2\pi \big\}.$$

Now with a function $x \in \mathcal{U}_{m_1,M}$, let $\Omega(x) \subset \mathbb{R}^2$ denote a domain being a subset of the unit circle, which is bounded by the curve $\gamma(x) = \{\zeta = (r,\theta) : \ r = x(\theta),\ 0 \le \theta \le 2\pi\}$. Consider the Laplacian field in $\Omega(x)$:

$$\begin{cases} \Delta u = 0 & \text{in} \quad \Omega(x), \\ u\,|_{\gamma(x)} = 0, & u\,|_{\Gamma} = \psi, \end{cases} \tag{45}$$

where $\psi \in C^3(\Gamma)$ is fixed and $\psi \ge 0$ does not vanish identically on Γ. Then there exists a unique classical solution $u(x) = u(x)(\zeta)$ to (45).

The inverse problem in this example is aimed at the identification of the interior subboundary $\gamma(x)$ from noisy data y^δ of $y := \frac{\partial u(x)}{\partial \nu}\,|_{\Gamma'}$ where Γ' is an arbitrary relatively open subset of Γ.

In the paper [20], a uniqueness assertion was proved, namely, that it can be concluded, for $x_1, x_2 \in \mathcal{U}_{m_1,M} \cap Q_{m_0}$, from the equality of two potential flux functions

$$\frac{\partial u(x_1)}{\partial \nu} = \frac{\partial u(x_2)}{\partial \nu} \quad \text{on} \ \ \Gamma'$$

that

$$x_1(\theta) = x_2(\theta), \qquad 0 \le \theta \le 2\pi.$$

Moreover, there exists a constant $C = C(m_0, m_1, M, \psi) > 0$ such that

$$\|x_1 - x_2\|_{C[0,2\pi]} \le \frac{C}{\left| \log \left\| \frac{\partial u(x_1)}{\partial \nu} - \frac{\partial u(x_2)}{\partial \nu} \right\|_{C^1(\Gamma')} \right|} \tag{46}$$

for all $x_1, x_2 \in \mathcal{U}_{m_1,M} \cap Q_{m_0}$.

The Banach spaces X and Y are fixed here as

$$X = C[0, 2\pi], \qquad Y = C^1(\Gamma'),$$

and the forward operator is introduced by the assignment

$$F(x) := \left.\frac{\partial u(x)}{\partial \nu}\right|_{\Gamma'}.$$

Taking into account the intrinsic ill-posedness of this inverse problem, nevertheless, it can be seen that the estimate (46) shows some weak, i.e., logarithmic, stability. This helps to overcome the ill-posedness here again if the admissible set is chosen suitably or the Tikhonov regularization is applied in an appropriate way.

The theory of the preceding sections applies if the functional Φ is chosen as

$$\Phi(x) = \left\| F(x) - y^\delta \right\|_Y^2 + \alpha \mathcal{R}(x)$$

with $\mathcal{R}(x)$ as a convex penalty term or if the penalty term is equipped with some stronger norm leading to

$$\Phi(x) = \|F(x) - y^\delta\|_Y^2 + \alpha \|x\|_{\tilde{X}}^2,$$

where $\tilde{X} = Q_{m_0} \cap Z$ and

$$Z = \left\{ x \in C^2[0, 2\pi] : \ \frac{d^k x}{d\theta^k}(0) = \frac{d^k x}{d\theta^k}(2\pi),\ k = 0, 1, 2 \right\}.$$

The conditional stability estimation (46) gives the convergence rate

$$\left\| x_\alpha^\delta - x^\dagger \right\|_{C[0,2\pi]} = \mathcal{O}\left(\frac{1}{|\log \delta|} \right) \quad \text{as} \quad \delta \to 0$$

for the parameter choice $\alpha = \delta^2$.

Similar inverse problems are discussed in the papers [14, 58]. The regularization methods outlined above can be used to treat those inverse problems, too.

Example III (Integral equation of the first kind with analytic kernel). Let D and D_1 be simple connected bounded domains in $\mathbb{R}^3$ such that $\overline{D} \cap \overline{D}_1 = \emptyset$. Consider an integral equation of the first kind:

$$[F(x)](\eta) := \int_D \frac{x(\zeta)}{|\eta - \zeta|^2} d\zeta = y(\eta), \qquad \eta \in D_1. \tag{47}$$

This type of integral equation is derived in the context of models for nondestructive testing (see [36]). In the original inverse problem, there is a nonlinear one. The integral equation (47), however, can be considered as a linearization of the original

problem. It was shown in [36] that the linearized problem (47) is close to the original problem under some assumptions on the size of domain D.

By $\overline{D} \cap \overline{D_1} = \emptyset$, the kernel $\frac{1}{|\eta-\zeta|^2}$ is analytic in $\eta \in D_1$ and $\zeta \in D$, so that (47) appears as a severely ill-posed linear operator equation.

In the paper [30], it was proved that if there are two functions $x_1, x_2 \in L^2(D)$ such that the corresponding y_1, y_2 satisfy

$$y_1(\eta) = y_2(\eta), \qquad \eta \in D_1,$$

then it holds

$$x_1(\zeta) = x_2(\zeta), \qquad \zeta \in D.$$

Moreover, the following conditional stability is proved: Fix $q > 3$ and

$$\mathcal{U}_M = \left\{ x \in W_0^{2,q}(D) : \ \|x\|_{W_0^{2,q}(D)} \le M \right\}.$$

Then there exists a constant $C = C(q, M, D, D_1) > 0$ such that

$$\|x\|_{L^2(D)} \le \frac{C}{\left|\log \|y\|_{H^1(D_1)}\right|} \tag{48}$$

for all $x \in \mathcal{U}_M$.

The linear forward operator F maps here from the space $X = L^2(D)$ to the space $Y = H^1(D_1)$. In spite of the original ill-posedness of the operator equation, the estimate (48) shows again logarithmic stability after appropriate restriction of the set of admissible solutions.

Variational regularization with the Tikhonov functional

$$\Phi(x) = \left\| F(x) - y^\delta \right\|_Y^2 + \alpha \|x - x^*\|_X^2$$

or alternatively with

$$\Phi(x) = \left\| F(x) - y^\delta \right\|_Y^2 + \alpha \|x\|_{\tilde{X}}^2$$

for $\tilde{X} = W_0^{2,q}(D)$ allows the application of the general theory to that example. In particular, the conditional stability estimation (48) yields the convergence rate

$$\|x_\alpha^\delta - x^\dagger\|_{L^2(D)} = \mathcal{O}\left(\frac{1}{|\log \delta|} \right) \quad \text{as} \quad \delta \to 0$$

whenever the a priori choice $\alpha = \delta^2$ of the regularization parameter is used.

Example IV (Identification of wave sources). Let $\Omega \subset \mathbb{R}^n$ be a bounded domain with C^2-boundary $\partial\Omega$. Consider

$$\begin{cases} \frac{\partial^2 u}{\partial t^2}(\zeta,t) = \Delta u(\zeta,t) + \lambda(t)x(\zeta), \quad x \in \Omega,\ 0 < t < T, \\ u(\zeta,0) = \frac{\partial u}{\partial t}(\zeta,0) = 0, \quad \zeta \in \Omega, \\ u(\zeta,t) = 0, \qquad \zeta \in \partial\Omega, \quad 0 < t < T. \end{cases} \tag{49}$$

Assume that

$$\lambda \in C^\infty[0,\infty), \quad \lambda(0) \neq 0, \tag{50}$$

and fix such λ. Then for any function $x \in L^2(\Omega)$, there exists a unique solution

$$u(x) \in C\left([0,T]; H^2(\Omega) \cap H_0^1(\Omega)\right) \cap C^1\left([0,T]; H_0^1(\Omega)\right) \cap C^2\left([0,T]; L^2(\Omega)\right).$$

The inverse problem under consideration here is the identification of $x = x(\zeta)$, $\zeta \in \Omega$, from observations y^δ of $y := \frac{\partial u(x)}{\partial \nu}|_{\partial\Omega\times(0,T)}$. Corresponding uniqueness and conditional stability results can be found in [121].

Let

$$\kappa \neq \frac{1}{4}, \frac{3}{4}, \qquad 0 \le \kappa \le 1,$$

and let $M > 0$ be arbitrarily given. Set

$$X_\kappa = \begin{cases} H^{2\kappa}(\Omega), & 0 \le \kappa < \frac{1}{4}, \\ H_0^{2\kappa}(\Omega), & \frac{1}{4} < \kappa \le 1,\ \kappa \neq \frac{3}{4}, \end{cases}$$

where $H^{2\kappa}(\Omega)$, $H_0^{2\kappa}(\Omega)$ denote the Sobolev spaces, and

$$\mathcal{U}_{M,\kappa} = \{x \in X_\kappa :\ \|x\|_{H^{2\kappa}(\Omega)} \le M\}.$$

Furthermore, assume

$$T > \operatorname{diam}\Omega \equiv \sup_{x,x' \in \Omega} |\zeta - \zeta'|. \tag{51}$$

Then, it is proved that there exists a constant $C = C(\Omega, T, \lambda, \kappa) > 0$ such that

$$\|x_1 - x_2\|_{L^2(\Omega)} \le C M^{\frac{1}{2\kappa+1}} \left\| \frac{\partial u(x_1)}{\partial \nu} - \frac{\partial u(x_2)}{\partial \nu} \right\|_{L^2(\partial\Omega\times(0,T))}^{\frac{2\kappa}{2\kappa+1}} \tag{52}$$

for all $x_1, x_2 \in \mathcal{U}_{M,\kappa}$.

The definition

$$F(x) = y := \left.\frac{\partial u(x)}{\partial \nu}\right|_{\partial\Omega\times(0,T)}$$

of the forward operator $F : X \to Y$ is well defined for the Hilbert spaces $X = L^2(\Omega)$ and $Y = L^2(\partial\Omega \times (0,T))$. In contrast to Example I, where also a wave equation is under consideration, F appears here as a linear operator with nonclosed range. However, the estimate (52) shows that this inverse problem possesses even Hölder type stability if we choose the admissible set suitably.

With respect to the regularization methods from the previous sections, the functional Φ can be chosen as

$$\Phi(x) = \left\| F(x) - y^\delta \right\|_Y^2 + \alpha \|x - x^*\|_X^2$$

or as

$$\Phi(x) = \left\| F(x) - y^\delta \right\|_Y^2 + \alpha \|x\|_{\tilde{X}}^2,$$

where $\tilde{X} = X_\kappa$. Then the conditional stability estimation (52) gives here the Hölder convergence rate

$$\left\| x_\alpha^\delta - x^\dagger \right\|_{L^2(\Omega)} = \mathcal{O}\left(\delta^{\frac{2\kappa}{2\kappa+1}}\right) \quad \text{as} \quad \delta \to 0$$

with the choice $\alpha = \delta^2$.

Example V (Identification of potential in an elliptic equation). Let Ω be a simply connected domain in $\mathbb{R}^3$ with the C^2-boundary $\partial\Omega$. Consider the following problem

$$\begin{cases} \Delta u + x \cdot u = 0, & \text{in} \quad \Omega \\ u = f, & \text{on} \quad \partial\Omega \end{cases} \tag{53}$$

with functions $x \in L^2(\Omega)$ and $f \in H^{\frac{1}{2}}(\partial\Omega)$.

Assuming that zero is not the Dirichlet eigenvalue of the Schrödinger operator $\Delta + x$ on the domain Ω; it is known that there exists unique solution $u \in H^1(\Omega)$ for this problem. Then the Dirichlet-to-Neumann map $\Lambda_x : H^{\frac{1}{2}}(\partial\Omega) \to H^{-\frac{1}{2}}(\partial\Omega)$ is defined as

$$\Lambda_x f = \left.\frac{\partial u}{\partial \nu}\right|_{\partial\Omega}, \tag{54}$$

where ν is the unit outer normal with respect to $\partial\Omega$.

The inverse problem under consideration here addresses the recovery of the not directly observable potential function $x(\zeta)$, for $\zeta \in \Omega$, from data y delivered by Λ_x. This is specified as follows: An infinite sequence $\{V_N\}_{N=1}^\infty$ of N-dimensional

subspaces $V_N = \text{span}(f_1, f_2, \ldots, f_N)$ is considered generated by a basis $\{f_j\}_{j=1}^{\infty}$ in $H^1(\partial\Omega)$, i.e.,

$$V_N \subset V_{N+1} \subset H^1(\partial\Omega) \qquad \text{and} \qquad \bigcup_{N=1}^{\infty} V_N \ \text{ is dense in } \ H^1(\partial\Omega).$$

In this context, it is assumed that the finite dimensional spaces V_N, $N = 1, 2, \ldots$, have the following properties:

1. For any $g \in H^1(\partial\Omega)$, there exists a $g_N \in V_N$ and a function $\beta(N)$, which satisfies $\lim_{N\to\infty} \beta(N) = 0$, such that

$$\|g - g_N\|_{H^{\frac{1}{2}}(\partial\Omega)} \leq \beta(N)\|g\|_{H^1(\partial\Omega)}. \tag{55}$$

2. There exists a constant $C > 0$, which is independent of g, such that

$$\|g_N\|_{H^{\frac{1}{2}}(\partial\Omega)} \leq C\|g\|_{H^{\frac{1}{2}}(\partial\Omega)}. \tag{56}$$

The following result is proved in [29]: Suppose that $x_j \in H^s(\Omega)$, $j = 1, 2$, with $s > \frac{3}{2}$, satisfy

$$\|x_j\|_{H^s(\Omega)} \leq M$$

for some constant $M > 0$. Then there exists a constant $C > 0$, which depends on M, such that

$$\|x_1 - x_2\|_{L^2(\Omega)} \leq C\omega\left(\|\Lambda_{x_1} - \Lambda_{x_2}\|_{V_N} + \beta(N)\right) \tag{57}$$

for N large enough and $\|\Lambda_{x_1} - \Lambda_{x_2}\|_{V_N}$ small enough. Precisely, it is here $\omega(t) = \left(\frac{1}{\log\frac{1}{t}}\right)^{\gamma}$ with some $0 < \gamma \leq 1$ taking into account that

$$\|\Lambda_{x_1} - \Lambda_{x_2}\|_{V_N} = \sup_{\phi\in V_N, \|\phi\|_{H^{\frac{1}{2}}(\partial\Omega)}=1} |\langle(\Lambda_{x_1} - \Lambda_{x_2})\phi, \phi\rangle|,$$

where $\langle\cdot,\cdot\rangle$ is the dual pairing between $H^{-\frac{1}{2}}(\partial\Omega)$ to $H^{\frac{1}{2}}(\partial\Omega)$.

Here the forward operator F is defined as

$$F(q) := \Lambda|_{Y_N}.$$

This is a nonlinear operator mapping from the space $X = L^2(\Omega)$ into the space $Y \in \mathcal{L}\left(L^2(\partial\Omega), H^{\frac{1}{2}}(\partial\Omega)\right)$, which represents the space of bounded linear operators mapping between $L^2(\partial\Omega)$ and $H^{\frac{1}{2}}(\partial\Omega)$. Moreover, consider the restriction $Y_N =$

$Y|_{V_N}$ of Y generated by the subspace V_N. The original problem of finding x from Λ_x data is ill-posed, but even without the uniqueness of the inverse problem, the estimate (57) shows some stability behavior of logarithmic type under the associated restrictions on the expected solution. Again for the Tikhonov regularization with functionals

$$\Phi(x) = \left\| F(x) - y^\delta \right\|_{Y_N}^2 + \alpha \|x - x^*\|_X^2$$

or

$$\Phi(x) = \left\| F(x) - y^\delta \right\|_{Y_N}^2 + \alpha \|x\|_{\tilde{X}}^2,$$

where $\tilde{X} = H^s$ with $s > \frac{3}{2}$, the theory of sections "Tikhonov Regularization in Hilbert Spaces with Quadratic Misfit and Penalty Terms" and "Variational Regularization in Banach Spaces with Convex Penalty Term" is applicable. From the latter section, with the conditional stability estimation (57), the convergence rate

$$\left\| x_\alpha^\delta - x^\dagger \right\|_{L^(\Omega)} = \mathcal{O}\left((\log(1/\delta))^{-\gamma} \right) \quad \text{as} \quad \delta \to 0$$

can be derived for the parameter choice $\alpha = \delta^2$.

Example VI (Inverse problems for autoconvolution equations). For the space $X = Y = L^2(0,1)$ of quadratically integrable real functions defined on the unit interval, consider as first variant of this example the autoconvolution equation $x * x = y$, where with reference to Eq. (1),

$$[F(x)](s) := \int_0^s x(s-t)x(t)dt, \quad 0 \le s \le 1 \tag{58}$$

is the corresponding forward operator with $\mathcal{D}(F) = X$. This operator equation of quadratic type occurs in physics of spectra, in optics, and in stochastics, often as part of a more complex task (see, e.g., [11, 70, 108]). A series of studies on deautoconvolution and regularization have been published for the setting (58); see, for example, [34, 100]. Some first basic mathematical analysis of the autoconvolution equation can already be found in the paper [48]. Moreover, a regularization approach for general quadratic operator equations was suggested in the recent paper [43]. Because of their weak nonlinearity, deautoconvolution problems are not seen as difficult, and hence, little attention is paid to them wrongly. However, there is a deficit in convergence rates for regularized solutions x_α^δ obtained by the classical form of Tikhonov regularization in Hilbert spaces as minimizers of the functional (5). It can be shown that Assumption 1 applies and that the inequality (7) from Theorem 2 is satisfied with $L = 2$ and for arbitrary large radii ρ, even as an equation

$$\left\| F(x) - F\left(x^\dagger\right) - F'\left(x^\dagger\right)\left(x - x^\dagger\right) \right\|_X = \left\| x - x^\dagger \right\|_X^2,$$

where

$$\left[F'(x)h\right](s) = 2\int_0^s x(s-t)h(t)dt, \quad 0 \le s \le 1, \quad h \in L^2(0,1),$$

characterizes the Fréchet derivative of F from (58) at the point $x^\dagger$. Note that it is very specific phenomenon here that the nonlinear operator F is not compact, but $F'(x^\dagger)$ is a compact linear operator mapping in $L^2(0,1)$. One consequence of this specific interplay between F and its derivative for the deautoconvolution problem from (58), which is locally ill-posed everywhere, is the fact that the classical convergence rate theory developed for the Tikhonov regularization of nonlinear ill-posed problems reaches its limits if standard source condition

$$x^\dagger - x^* = F'\left(x^\dagger\right)^* w, \qquad w \in Y,$$

fails. Please refer to [24] for details. On the other hand, convergence rate results based on Hölder source conditions with small Hölder exponent and logarithmic source conditions or on the method of approximate source conditions (cf. [58]) are not applicable since qualified nonlinearity conditions like (36) cannot be proven according to current knowledge.

For a function $x : \mathbb{R} \to \mathbb{R}$ with support on $[0,1]$, the autoconvolution $x * x$ is a function with support on $[0,2]$. Hence, the strength of ill-posedness for the deautoconvolution problem according to the forward operator (58) can be reduced if observations of $[F(x)](s)$ for all $0 \le s \le 2$ are taken into account (full data case). Please refer to [46] for details and numerical case studies. In a second variant of this example with applications in laser optics, the full data case is exploited, but for complex functions and with an additional kernel. This variety considers a generalized autoconvolution equation motivated by problems of ultrashort laser pulse characterization arising in the context of the self-diffraction SPIDER method, and the reader is referred to the recent paper [47] for physical details and the experimental setting of this problem. In this variant, the focus is on a kernel-based, complex-valued, and full data analog to (58). Take into account $L^2_{\mathbb{C}}$-spaces of quadratically integrable complex-valued functions over finite real intervals set for the corresponding normed spaces $X = L^2_{\mathbb{C}}(0,1)$ and $Y = L^2_{\mathbb{C}}(0,1)$ and consider for the associated operator equation (1) the nonlinear forward operator

$$[F(x)](s) := \begin{cases} \int_0^s k(s,t)x(s-t)x(t)dt & \text{if } 0 \le s \le 1 \\ \int_{s-1}^1 k(s,t)x(s-t)x(t)dt & \text{if } 1 < s \le 2 \end{cases} \tag{59}$$

mapping from $L^2_{\mathbb{C}}(0,1)$ to $L^2_{\mathbb{C}}(0,2)$ with domain $\mathcal{D}(F) = L^2_{\mathbb{C}}(0,1)$. Every function $x \in L^2_{\mathbb{C}}(0,1)$ can be represented as $x(t) = A(t)\, e^{i\phi(t)}$, $0 \le t \le 1$, with the nonnegative amplitude (modulus) function $A = |x|$ and the phase function $\phi : [0,1] \to \mathbb{R}$. For the SPIDER technology, in particular, the phase function is to be determined from noisy observations of the complex function y, whereas information about the amplitude function can be verified by alternative measurements. In [47] some mathematical studies and a regularization approach for this specific problem have been presented, and further analytic investigations for the specific case of a constant kernel k can be found in [24].

4 Conclusion

This chapter has presented some theoretic results including convergence and convergence rates assertions on direct regularization methods for nonlinear inverse problems formulated in the setting of infinite dimensional Hilbert or Banach spaces. The inverse problems can be written as ill-posed nonlinear operator equations with the consequence that their solutions tend to be unstable with respect to data perturbations. To overcome that drawback, regularization methods use stable auxiliary problems, which are close to the original inverse problem. A regularization parameter controls the trade-off between approximation and stability. For direct regularization methods, the auxiliary problems are mostly minimization problems in abstract spaces, where a weighted sum of a residual term that expresses the data misfit and a stabilizing penalty term expressing expected solution properties has to be minimized. In this context the regularization parameter controls the relative weight of both terms. Furthermore, six examples are given that show the wide range of applicability for such regularization methods in the light of specific inverse problems. More than 120 references at the end of this chapter survey the relevant literature in this field.

Cross-References

- Iterative Solution Methods
- Linear Inverse Problems
- Numerical Methods and Applications in Total Variation Image Restoration
- Statistical Methods in Imaging
- Total Variation in Imaging

References

1. Acar, R., Vogel, C.R.: Analysis of bounded variation penalty methods for ill-posed problems. Inverse Probl. **10**(6), 1217–1229 (1994)

2. Ammari, H.: An Introduction to Mathematics of Emerging Biomedical Imaging. Springer, Berlin (2008)
3. Anzengruber, S.W., Hofmann, B., Mathé, P.: Regularization properties of the discrepancy principle for Tikhonov regularization in Banach spaces. Appl. Anal. (2013). http://dx.doi.org/10.1080/00036811.2013.833326.
4. Anzengruber, S.W., Hofmann, B., Ramlau, R.: On the interplay of basis smoothness and specific range conditions occurring in sparsity regularization. Inverse Probl. **29**(12), 125002(21pp) (2013)
5. Anzengruber, S.W., Ramlau, R.: Morozov's discrepancy principle for Tikhonov-type functionals with nonlinear operators. Inverse Probl. **26**(2), 025001(17pp) (2010)
6. Anzengruber, S.W., Ramlau, R.: Convergence rates for Morozov's discrepancy principle using variational inequalities. Inverse Probl. **27**(10), 105007(18pp) (2011)
7. Bakushinsky, A., Goncharsky, A.: Ill-Posed Problems: Theory and Applications. Kluwer, Dordrecht (1994)
8. Bakushinsky, A.B., Kokurin, M.Yu.: Iterative Methods for Approximate Solution of Inverse Problems. Springer, Dordrecht (2004)
9. Banks, H.T., Kunisch, K.: Estimation Techniques for Distributed Parameter Systems. Birkhäuser, Boston, (1989)
10. Baumeister, J.: Stable Solution of Inverse Problems. Vieweg, Braunschweig (1987)
11. Baumeister, J.: Deconvolution of appearance potential spectra. In: Kleinman R., Kress R., Martensen E. (eds.) Direct and Inverse Boundary Value Problems. Methoden und Verfahren der mathematischen Physik, vol. 37, pp. 1–13. Peter Lang, Frankfurt am Main (1991)
12. Benning, M., Burger, M.: Error estimates for general fidelities. Electron. Trans. Numer. Anal. **38**, 44–68 (2011)
13. Benning, M., Burger, M.: Ground states and singular vectors of convex variational regularization methods. arXiv:1211.2057v1 (2012)
14. Beretta, E., Vessella, S.: Stable determination of boundaries from Cauchy data. SIAM J. Math. Anal. **30**, 220–232 (1999)
15. Bertero, M., Boccacci, P.: Introduction to Inverse Problems in Imaging. Institute of Physics Publishing, Bristol (1998)
16. Bonesky, T., Kazimierski, K., Maass, P., Schöpfer, F., Schuster, T.: Minimization of Tikhonov functionals in Banach spaces. Abstr. Appl. Anal. Art. ID 192679, 19pp (2008)
17. Boţ, R.I., Hofmann, B.: An extension of the variational inequality approach for obtaining convergence rates in regularization of nonlinear ill-posed problems. J. Integral Equ. Appl. **22**(3), 369–392 (2010)
18. Boţ, R.I., Hofmann, B.: The impact of a curious type of smoothness conditions on convergence rates in ℓ^1-regularization. Eurasian J. Math. Comput. Appl. **1**(1), 29–40 (2013)
19. Bredies, K., Lorenz, D.A.: Regularization with non-convex separable constraints. Inverse Probl. **25**(8), 085011(14pp) (2009)
20. Bukhgeim, A.L., Cheng, J., Yamamoto, M.: Stability for an inverse boundary problem of determining a part of a boundary. Inverse Probl **15**, 1021–1032 (1999)
21. Burger, M., Flemming, J., Hofmann, B.: Convergence rates in ℓ^1-regularization if the sparsity assumption fails. Inverse Probl. **29**(2), 025013(16pp) (2013)
22. Burger, M., Osher, S.: Convergence rates of convex variational regularization. Inverse Probl. **20**(5), 1411–1421 (2004)
23. Burger, M., Resmerita, E., He, L.: Error estimation for Bregman iterations and inverse scale space methods in image restoration. Computing **81**(2–3), 109–135 (2007)
24. Bürger, S., Hofmann, B.: About a deficit in low order convergence rates on the example of autoconvolution. Appl. Anal. (2014, to appear). Preprint 2013–17, Preprintreihe der Fakultät für Mathematik der TU Chemnitz, 2013. http://nbn-resolving.de/urn:nbn:de:bsz:ch1-qucosa-130630.
25. Chavent, G.: Nonlinear Least Squares for Inverse Problems. Springer, Dordrecht (2009)
26. Chavent, G., Kunisch, K.: On weakly nonlinear inverse problems. SIAM J. Appl. Math. **56**(2), 542–572 (1996)

27. Chavent, G., Kunisch, K.: State space regularization: geometric theory. Appl. Math. Opt. **37**(3), 243–267 (1998)
28. Cheng, J., Hofmann, B., Lu, S.: The index function and Tikhonov regularization for ill-posed problems. J. Comput. Appl. Math. (2013). http://dx.doi.org/10.1016/j.cam.2013.09.035.
29. Cheng, J., Nakamura, G.: Stability for the inverse potential problem by finite measurements on the boundary. Inverse Probl. **17**, 273–280 (2001)
30. Cheng, J., Yamamoto, M.: Conditional stabilizing estimation for an integral equation of first kind with analytic kernel. J. Integral Equ. Appl. **12**, 39–61 (2000)
31. Cheng, J., Yamamoto, M.: One new strategy for a priori choice of regularizing parameters in Tikhonov's regularization. Inverse Probl. **16**(4), L31–L38 (2000)
32. Clason, C.: L^∞ fitting for inverse problems with uniform noise. Inverse Probl. **28**(10), 104007(18pp) (2012)
33. Colton, D., Kress, R.: Inverse Acoustic and Electromagnetic Scattering Theory, 3rd edn. Springer, New York (2013)
34. Dai, Z., Lamm, P.K.: Local regularization for the nonlinear inverse autoconvolution problem. SIAM J. Numer. Anal. **46**(2), 832–868 (2008)
35. Engl, H.W., Hanke, M., Neubauer, A.: Regularization of Inverse Problems. Kluwer, Dordrecht (1996, 2000)
36. Engl, H.W., Isakov, V.: On the identifiability of steel reinforcement bars in concrete from magnetostatic measurements. Eur. J. Appl. Math. **3**, 255–262 (1992)
37. Engl, H.W., Kunisch, K., Neubauer, A.: Convergence rates for Tikhonov regularisation of non-linear ill-posed problems. Inverse Probl. **5**(4), 523–540 (1989)
38. Engl, H.W., Zou, J.: A new approach to convergence rate analysis of Tikhonov regularization for parameter identification in heat conduction. Inverse Probl. **16**(6), 1907–1923 (2000)
39. Favaro, P., Soatto, S. 3-D Shape Estimation and Image Restoration. Exploiting Defocus and Motion Blur. Springer, London (2007)
40. Fischer, B., Modersitzki, J.: Ill-posed medicine – an introduction to image registration. Inverse Probl. **24**(3), 034008(16pp) (2008)
41. Flemming, J.: Generalized Tikhonov Regularization and Modern Convergence Rate Theory in Banach Spaces. Shaker Verlag, Aachen (2012)
42. Flemming, J.: Variational smoothness assumptions in convergence rate theory – an overview. J. Inverse Ill-Posed Probl. **21**(3), 395–409 (2013)
43. Flemming, J.: Regularization of autoconvolution and other ill-posed quadratic equations by decomposition. J. Inverse and Ill-Posed Probl. **22** (2014). doi:10.1515/jip-2013-0038
44. Flemming, J., Hofmann, B.: A new approach to source conditions in regularization with general residual term. Numer. Funct. Anal. Optim. **31**(3), 254–284 (2010)
45. Flemming, J., Hofmann, B.: Convergence rates in constrained Tikhonov regularization: equivalence of projected source conditions and variational inequalities. Inverse Probl. **27**(8), 085001(11pp) (2011)
46. Fleischer, G., Hofmann, B.: On inversion rates for the autoconvolution equation. Inverse Probl. **12**(4), 419–435 (1996)
47. Gerth, D., Hofmann, B., Birkholz, S., Koke, S., Steinmeyer, G.: Regularization of an autoconvolution problem in ultrashort laser pulse characterization. Inverse Probl. Sci. Eng. **22**(2), 245–266 (2014)
48. Gorenflo, R., Hofmann, B.: On autoconvolution and regularization. Inverse Probl. **10**(2), 353–373 (1994)
49. Grasmair, M.: Well-posedness and convergence rates for sparse regularization with sublinear l^q penalty term. Inverse Probl. Imaging **3**(3), 383–387 (2009)
50. Grasmair, M.: Generalized Bregman distances and convergence rates for non-convex regularization methods. Inverse Probl. **26**(11), 115014(16pp) (2010)
51. Grasmair, M.: Variational inequalities and higher order convergence rates for Tikhonov regularisation on Banach spaces. J. Inverse Ill-Posed Probl. **21**(3), 379–394 (2013)
52. Grasmair, M., Haltmeier, M., Scherzer, O.: Sparse regularization with ℓ^q penalty term. Inverse Probl. **24**(5), 055020(13pp) (2008)

53. Grasmair, M., Haltmeier, M., Scherzer, O.: The residual method for regularizing ill-posed problems. Appl. Math. Comput. **218**(6), 2693–2710 (2011)
54. Groetsch, C.W.: The Theory of Tikhonov Regularization for Fredholm Integral Equations of the First Kind. Pitman, Boston (1984)
55. Hansen, P.C.: Rank-Deficient and Discrete Ill-Posed Problems. SIAM, Philadelphia (1998)
56. Hao, D.N., Quyen, T.N.T.: Convergence rates for Tikhonov regularization of coefficient identification problems in Laplace-type equations. Inverse Probl. **26**(12), 125014(23pp) (2010)
57. Hein, T.: Tikhonov regularization in Banach spaces – improved convergence rates results. Inverse Probl. **25**(3), 035002(18pp) (2009)
58. Hein, T., Hofmann, B.: Approximate source conditions for nonlinear ill-posed problems – chances and limitations. Inverse Probl. **25**(3), 035003(16pp) (2009)
59. Hofmann, B.: Regularization for Applied Inverse and Ill-Posed Problems. Teubner, Leipzig (1986)
60. Hofmann, B., Kaltenbacher, B., Pöschl, C., Scherzer, O.: A convergence rates result for Tikhonov regularization in Banach spaces with non-smooth operators. Inverse Probl. **23**(3), 987–1010 (2007)
61. Hofmann, B., Mathé, P.: Analysis of profile functions for general linear regularization methods. SIAM J. Numer. Anal. **45**(3), 1122–1141 (2007)
62. Hofmann, B., Mathé, P.: Parameter choice in Banach space regularization under variational inequalities. Inverse Probl. **28**(10), 104006(17pp) (2012)
63. Hofmann, B., Mathé, P., Pereverzev, S.V.: Regularization by projection: approximation theoretic aspects and distance functions. J. Inverse Ill-Posed Probl. **15**(5), 527–545 (2007)
64. Hofmann, B., Scherzer, O.: Factors influencing the ill-posedness of nonlinear problems. Inverse Probl. **10**(6), 1277–1297 (1994)
65. Hofmann, B., Yamamoto, M.: On the interplay of source conditions and variational inequalities for nonlinear ill-posed problems. Appl. Anal. **89**(11), 1705–1727 (2010)
66. Hohage, T., Pricop, M.: Nonlinear Tikhonov regularization in Hilbert scales for inverse boundary value problems with random noise. Inverse Probl. Imaging **2**(2), 271–290 (2008)
67. Imanuvilov, O.Yu., Yamamoto, M.: Global uniqueness and stability in determining coefficients of wave equations. Commun. Partial Differ. Equ. **26**, 1409–1425 (2001)
68. Isakov, V.: Inverse Problems for Partial Differential Equations. Springer, New York (2006)
69. Ito, K., Kunisch, K.: On the choice of the regularization parameter in nonlinear inverse problems. SIAM J. Optim. **2**(3), 376–404 (1992)
70. Janno, J., Wolfersdorf, L.v.: A general class of autoconvolution equations of the third kind. Z. Anal. Anwendungen **24**(3), 523–543 (2005)
71. Jiang, D., Feng, H., Zou, J.: Convergence rates of Tikhonov regularizations for parameter identification in a parabolic-elliptic system. Inverse Probl. **28**(10), 104002(20pp) (2012)
72. Jin, B., Zou, J.: Augmented Tikhonov regularization. Inverse Probl. **25**(2), 025001(25pp) (2009)
73. Kabanikhin, S.I.: Inverse and Ill-Posed Problems – Theory and Applications. Inverse and Ill-Posed Problems Series, vol. 55. Walter de Gruyter, Berlin (2011)
74. Kaipio, J., Somersalo, E.: Statistical and Computational Inverse Problems. Springer, New York (2005)
75. Kaltenbacher, B.: A note on logarithmic convergence rates for nonlinear Tikhonov regularization. J. Inverse Ill-Posed Probl. **16**(1), 79–88 (2008)
76. Kaltenbacher, B., Neubauer, A., Scherzer, O.: Iterative Regularization Methods for Nonlinear Ill-Posed Problems. Walter de Gruyter, Berlin (2008)
77. Kirsch, A.: An Introduction to the Mathematical Theory of Inverse Problems, 2nd edn. Springer, New York (2011)
78. Klann, E., Kuhn, M., Lorenz, D.A., Maass, P., Thiele, H.: Shrinkage versus deconvolution. Inverse Probl. **23**(5), 2231–2248 (2007)
79. Kress, R.: Linear Integral Equations, 2nd edn. Springer, New York (1999)

80. Lamm, P.K., Dai, Z.: On local regularization methods for linear Volterra equations and nonlinear equations of Hammerstein type. Inverse Probl. **21**(5), 1773–1790 (2005)
81. Lattès, R., Lions, J.-L.: The Method of Quasi-Reversibility. Applications to Partial Differential Equations. Modern Analytic and Computational Methods in Science and Mathematics, vol. 18. American Elsevier, New York (1969)
82. Lorenz, D., Rösch, A.: Error estimates for joint Tikhonov and Lavrentiev regularization of constrained control problems. Appl. Anal. **89**(11), 1679–1691 (2010)
83. Lorenz, D., Worliczek, N.: Necessary conditions for variational regularization schemes. Inverse Probl. **29**(7), 075016(19pp), (2013)
84. Louis, A.K.: Inverse und schlecht gestellte Probleme. Teubner, Stuttgart (1989)
85. Louis, A.K.: Approximate inverse for linear and some nonlinear problems. Inverse Probl. **11**(6), 1211–1223 (1995)
86. Liu, F., Nashed, M.Z.: Regularization of nonlinear ill-posed variational inequalities and convergence rates. Set-Valued Anal. **6**(4), 313–344 (1998)
87. Lu, S., Flemming, J.: Convergence rate analysis of Tikhonov regularization for nonlinear ill-posed problems with noisy operators. Inverse Probl. **28**(10), 104003(19pp) (2012)
88. Lu, S., Pereverzev, S.V., Ramlau, R.: An analysis of Tikhonov regularization for nonlinear ill-posed problems under a general smoothness assumption. Inverse Probl. **23**(1), 217–230 (2007)
89. Mahale, P., Nair, M.T.: Tikhonov regularization of nonlinear ill-posed equations under general source conditions. J. Inverse Ill-Posed Probl. **15**(8), 813–829 (2007)
90. Mathé, P., Pereverzev, S.V.: Geometry of linear ill-posed problems in variable Hilbert scales. Inverse Probl. **19**(3), 789–803 (2003)
91. Modersitzki, J.: FAIR. Flexible Algorithms for Image Registration. SIAM, Philadelphia (2009)
92. Morozov, V.A.: Methods for Solving Incorrectly Posed Problems. Springer, New York (1984)
93. Natterer, F.: Imaging and inverse problems of partial differential equations. Jahresber. Dtsch. Math.-Ver. **109**(1), 31–48 (2007)
94. Natterer, F., Wübbeling, F.: Mathematical Methods in Image Reconstruction. SIAM, Philadelphia (2001)
95. Neubauer, A.: Tikhonov regularization for nonlinear ill-posed problems: optimal convergence rate and finite dimensional approximation. Inverse Probl. **5**(4), 541–558 (1989)
96. Neubauer, A.: On enhanced convergence rates for Tikhonov regularization of nonlinear ill-posed problems in Banach spaces. Inverse Probl. **25**(6), 065009(10pp) (2009)
97. Neubauer, A., Hein, T., Hofmann, B., Kindermann, S., Tautenhahn, U.: Improved and extended results for enhanced convergence rates of Tikhonov regularization in Banach spaces. Appl. Anal. **89**(11), 1729–1743 (2010)
98. Phillips, D.L.: A technique for the numerical solution of certain integral equations of the first kind. J. ACM **9**(1), 84–97 (1962)
99. Pöschl, C.: Tikhonov Regularization with General Residual Term. PhD thesis, University of Innsbruck, Austria, (2008)
100. Ramlau, R.: Morozov's discrepancy principle for Tikhonov-regularization of nonlinear operators. Numer. Funct. Anal. and Optim. **23**(1–2), 147–172 (2002)
101. Ramlau, R.: TIGRA – an iterative algorithm for regularizing nonlinear ill-posed problems. Inverse Probl. **19**(2), 433–465 (2003)
102. Resmerita, E.: Regularization of ill-posed problems in Banach spaces: convergence rates. Inverse Probl. **21**(4), 1303–1314 (2005)
103. Resmerita, E., Scherzer, O.: Error estimates for non-quadratic regularization and the relation to enhancement. Inverse Probl. **22**(3), 801–814 (2006)
104. Rondi, L.: Uniqueness and stability for the determination of boundary defects by electrostatic measurements. Proc. R. Soc. Edinb. Sect. A **130**, 1119–1151 (2000)
105. Scherzer, O., Engl, H.W., Kunisch, K.: Optimal a posteriori parameter choice for Tikhonov regularization for solving nonlinear ill-posed problems. SIAM J. Numer. Anal. **30**(6), 1796–1838 (1993)

106. Scherzer, O. (ed.): Mathematical Models for Registration and Applications to Medical Imaging. Mathematics in Industry 10. The European Consortium for Mathematics in Industry. Springer, Berlin (2006)
107. Scherzer, O., Grasmair, M., Grossauer, H., Haltmeiner, M., Lenzen, F.: Variational Methods in Imaging. Springer, New York (2009)
108. Schleicher, K.-Th., Schulz, S.W., Gmeiner, R., Chun, H.-U.: A computational method for the evaluation of highly resolved DOS functions from APS measurements. J. Electron Spectrosc. Relat. Phenom. **31**, 33–56 (1983)
109. Schuster, T., Kaltenbacher, B., Hofmann, B., Kazimierski, K.S.: Regularization Methods in Banach Spaces. Radon Series on Computational and Applied Mathematics, vol. 10. Walter de Gruyter, Berlin/Boston (2012)
110. Seidman, T.I., Vogel, C.R.: Well posedness and convergence of some regularization methods for nonlinear ill posed problems. Inverse Probl. **5**(2), 227–238 (1989)
111. Tautenhahn, U.: On a general regularization scheme for nonlinear ill-posed problems. Inverse Probl. **13**(5), 1427–1437 (1997)
112. Tautenhahn, U.: On the method of Lavrentiev regularization for nonlinear ill-posed problems. Inverse Probl. **18**(1), 191–207 (2002)
113. Tautenhahn, U., Jin, Q.: Tikhonov regularization and a posteriori rules for solving nonlinear ill-posed problems. Inverse Probl. **19**(1), 1–21 (2003)
114. Tikhonov, A.N.: Solution of incorrectly formulated problems and the regularization method. Dokl. Akad. Nauk SSR **151**, 501–504 (1963)
115. Tikhonov, A.N., Arsenin, V.Y.: Solutions of Ill-Posed Problems. Wiley, New York (1977)
116. Tikhonov, A.N., Goncharsky, A.V., Stepanov, V.V., Yagola, A.G.: Numerical Methods for the Solution of Ill-Posed Problems. Kluwer, Dordrecht (1995)
117. Tikhonov, A.N., Leonov, A.S., Yagola, A.G.: Nonlinear Ill-Posed Problems, vols. 1 and 2. Series Applied Mathematics and Mathematical Computation, vol. 14. Chapman & Hall, London (1998)
118. Vasin, V.V.: Some tendencies in the Tikhonov regularization of ill-posed problems. J. Inverse Ill-Posed Probl. **14**(8), 813–840 (2006)
119. Vogel, C.: Computational Methods for Inverse Problems. SIAM, Philadelphia (2002)
120. Werner, F., Hohage, T.: Convergence rates in expectation for Tikhonov-type regularization of inverse problems with Poisson data. Inverse Probl. **28**(10), 104004(15pp) (2012)
121. Yamamoto, M.: On ill-posedness and a Tikhonov regularization for a multidimensional inverse hyperbolic problem. J. Math. Kyoto Univ. **36**, 825–856 (1996)
122. Zarzer, C.A.: On Tikhonov regularization with non-convex sparsity constraints. Inverse Probl. **25**(2), 025006(13pp) (2009)

Distance Measures and Applications to Multimodal Variational Imaging

Christiane Pöschl and Otmar Scherzer

Contents

1 Introduction 126
2 Distance Measures 127
Deterministic Pixel Measure 127
Morphological Measures 128
Statistical Distance Measures 130
Statistical Distance Measures (Density Based) 132
f-Information 141
Distance Measures Including Statistical Prior Information 145
3 Mathematical Models for Variational Imaging 146
4 Registration 148
5 Recommended Reading 151
6 Conclusion 152
Cross-References 153
References 153

Abstract

Today *imaging* is rapidly improving by increased specificity and sensitivity of measurement devices. However, even more diagnostic information can be gained by combination of data recorded with different imaging systems.

C. Pöschl (✉)
Institute of Mathematics, Alpen Adria Universität Klagenfurt, Klagenfurt, Austria
e-mail: christiane.poeschl@aau.at

O. Scherzer
Computational Science Center, University of Vienna, Vienna, Austria

RICAM, Austrian Academy of Sciences, Linz, Austria
e-mail: otmar.scherzer@univie.ac.at

© Springer Science+Business Media New York 2015

O. Scherzer (ed.), *Handbook of Mathematical Methods in Imaging*,
DOI 10.1007/978-1-4939-0790-8_4

1 Introduction

Today *imaging* is rapidly improving by increased specificity and sensitivity of measurement devices. However, even more diagnostic information can be gained by combination of data recorded with different imaging systems.

In particular in medicine, information of different modalities is used for diagnosis. From the various imaging technologies used in medicine, we mention exemplary *positron emission tomography* (PET), *single photon emission computed tomography* (SPECT), *magnetic resonance imaging* (MRI), *magnetic resonance spectroscopy* (MRS), X-ray, and *ultrasound*. Soft tissue can be well visualized in magnetic resonance scans, while bone structures are more easily discernible by X-ray imaging.

Image registration is an appropriate tool to align the information gained from different modalities. Thereby, it is necessary to use similarity measures that are able to compare images of different modalities, such that in a post-processing step the data can be fused and relevant information can be aligned.

The main challenge for computer-assisted comparison of images from different modalities is to define an appropriate distance measure between the images from different modalities.

Similarity measures of images can be categorized as follows:

1. Pixel-wise comparison of intensities.
2. A morphological measure defines the distance between images by the distance between their level sets.
3. Measures based on the image's gray value distributions.

In the following, we review distance measures for images according to the above catalog.

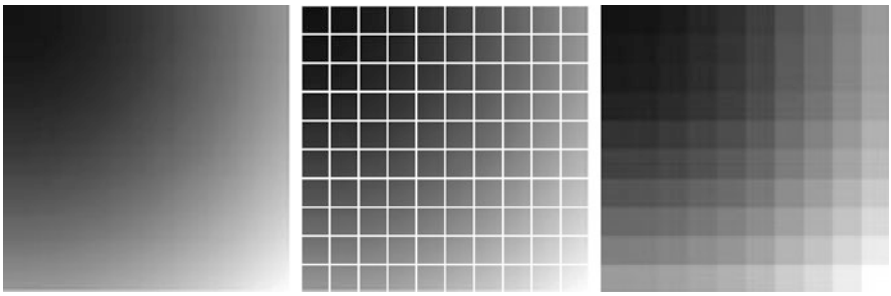
We use the notation Ω for the squared domain $(0,\ 1)^2$. Images are simultaneously considered as matrices or functions on Ω: A *discrete image* is an $N \times N$-matrix $U \in \{0,\ \ldots,\ 255\}^{N\times N}$. Each of the entries of the matrix represents an intensity value at a pixel. Therewith is associated a piecewise constant function

$$u_N(x) = \sum\nolimits_{i=1}^{N} \sum\nolimits_{j=1}^{N} U^{ij} \chi_{\Omega_{ij}}(x), \tag{1}$$

where

$$\Omega_{ij} := \left(\frac{i-1}{N}, \frac{i}{N}\right] \times \left(\frac{j-1}{N}, \frac{j}{N}\right] \text{ for } 1 \le i, j \le N,$$

and $\chi_{\Omega ij}$ is the characteristic function of Ω_{ij}. In the context of image processing, U^{ij} denotes the *pixel intensity* at the *pixel* $\chi_{\Omega ij}$. A *continuous* image is a function $u:\ \Omega \to \mathbb{R}$.



We emphasize that the measures for comparing images, presented below, can be applied in a straightforward way to higher-dimensional domains, for example, voxel data. However, here, for the sake of simplicity of notation and readability, we restrict attention to a two-dimensional squared domain Ω. Even more, we restrict attention to intensity data and do not consider vector-valued data, such as color images or tensor data. By this restriction, we exclude, for instance, feature-based intensity measures.

2 Distance Measures

In the following, we review distance measures for comparing discrete and continuous images. We review the standard and a morphological distance measure; both of them are deterministic. Moreover, based on the idea to consider images as random variable, we consider in the last two subsections two statistical approaches.

Deterministic Pixel Measure

The most widely used distance measures for discrete and continuous images are the l^p, L^p distance measures, respectively, in particular $p = 2$; see, for instance, the chapter ▸ Linear Inverse Problems in this handbook. There, two discrete images U_1 and U_2 are similar, if

$$\begin{aligned} \|U_1 - U_2\|_p &:= \left(\sum_{i=1}^N \sum_{j=1}^N \left|U_1^{ij} - U_2^{ij}\right|^p\right)^{\frac{1}{p}}, \quad 1 \le p < \infty, \\ \|U_2 - U_2\|_\infty &:= \sup_{i,j=1,\dots,N} \left|U_1^{ij} - U_2^{ij}\right|, \quad p = \infty, \end{aligned}$$

respectively, is small. Two continuous images $u_1,\ u_2 :\ \Omega \to \mathbb{R}$ are similar if

$$\begin{aligned} \|u_1 - u_2\|_p &:= \left(\int_\Omega |u_1(x) - u_2(x)|^p, dx\right)^{\frac{1}{p}} \quad 1 \le p < \infty, \\ \|u_2 - u_2\|_\infty &:= \text{ess sup}_{x,y} |u_1(x) - u_2(x)|, \quad p = \infty, \end{aligned}$$

is small. Here, ess sup denotes the essential supremum.

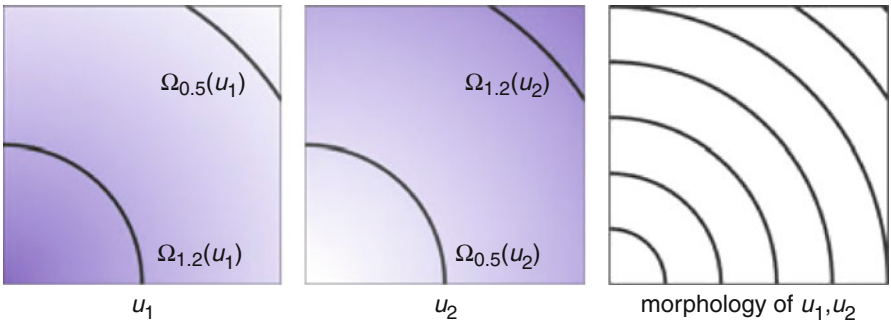


Fig. 1 The gray values of the images are completely different, but the images u_1, u_2 have the same morphology

Morphological Measures

In this subsection, we consider continuous images $u_i : \Omega \to [0, 255]$, $i = 1, 2$. u_1 and u_2 are *morphologically equivalent* (Fig. 1), if there exists a one-to-one gray value transformation $\beta : [0, 255] \to [0, 255]$, such that

$$\beta \circ u_1 = u_2.$$

Level sets of a continuous function u are defined as

$$\Omega_t(u) := \{x \in \Omega : u(x) = t\}.$$

The level sets $\Omega_{\mathbb{R}}(u) := \{\Omega_t(u) : t \in [0, 255]\}$ form the objects of an image that remain invariant under gray value transformations. The *normal field (Gauss map)* is given by the normals to the level lines and can be written as

$$\begin{array}{rcl} \mathbf{n}(u) : \Omega & \to & \mathbb{R}^d \\ x & \mapsto & \begin{cases} 0 & \text{if } \nabla u(x) = 0 \\ \frac{\nabla u(x)}{\|\nabla u(x)\|} & \text{else.} \end{cases} \end{array}$$

Droske and Rumpf [7] consider images as similar, if intensity changes occur at the same locations. Therefore, they compare the normal fields of the images with the similarity measure

$$\mathcal{S}_g(u_1, u_2) = \int_\Omega g(\mathbf{n}(u_1)(x), \mathbf{n}(u_2)(x))\, dx, \tag{2}$$

where they choose the function $g : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R} \geq 0$ appropriately. The vectors $\mathbf{n}(u_1)(x)$, $\mathbf{n}(u_2)(x)$ form an angle that is minimal if the images are morphologically equivalent. Therefore, an appropriate choice of the function g is an increasing

function of the minimal angle between v_1, v_2, and v_1, $-v_2$. For instance, setting g to be the cross or the negative dot product, we obtain

$$\begin{aligned}\mathcal{S}_\times(u_1,u_2) &= \tfrac{1}{2}\int_\Omega |\mathbf{n}(u_1)(x)\times\mathbf{n}(u_2)(x)|^2\,dx\\ \mathcal{S}_\circ(u_1,u_2) &= \tfrac{1}{2}\int_\Omega (1-\mathbf{n}(u_1)(x)\cdot\mathbf{n}(u_2)(x))^2\,dx.\end{aligned}$$

(The vectors $\mathbf{n}$ have to be embedded in $\mathbb{R}^3$ in order to calculate the cross product.)

Example 1. Consider the following scaled images $u_i : [0,\ 1]^2 \to [0,\ 1]$,

$$u_1(x) = x_1x_2,\ \ u_2(x) = 1-x_1x_2,\ \ u_3(x) = (1-x_1)x_2,$$

with gradients

$$\nabla u_1(x) = \begin{pmatrix} x_2\\ x_1\end{pmatrix},\quad \nabla u_2(x) = \begin{pmatrix} -x_2\\ -x_1\end{pmatrix},\quad \nabla u_3(x) = \begin{pmatrix} -x_2\\ 1-x_1\end{pmatrix}.$$

With $g(u,v) := \frac{1}{2}|u_1v_2 - u_2v_1|$, the functional $\mathcal{S}_g$ defined in (2) attains the following values for the particular images:

$$\begin{aligned}\mathcal{S}_g(u_1,u_2) &= \tfrac{1}{2}\int_\Omega |-x_2x_1 + x_2x_1|\,dx = 0\\ \mathcal{S}_g(u_2,u_3) &= \tfrac{1}{2}\int_\Omega |x_2x_1 + x_2x_1|\,dx = \tfrac{1}{4}\\ \mathcal{S}_g(u_3,u_1) &= \tfrac{1}{2}\int_\Omega |-x_2x_1 - (1-x_1)x_2|\,dx = \tfrac{1}{4}.\end{aligned}$$

The similarity measure indicates that u_1 and u_2 are morphologically identical.

The normalized gradient field is set valued in regions where the function is constant. Therefore, the numerical evaluation of the gradient field is highly unstable. To overcome this drawback, Haber and Modersitzki [15] suggested to use regularized normal gradient fields:

$$\begin{aligned}\mathbf{n}_\epsilon(u):\quad \Omega &\to \mathbb{R}^d\\ x &\mapsto \tfrac{\nabla u(x)}{\|\nabla u(x)\|_\epsilon}\end{aligned}$$

where $\|v\|_\epsilon := \sqrt{v^Tv+\epsilon^2}$ for every $v\in\mathbb{R}^d$. The parameter ϵ is connected to the estimated noise level in the image. In regions where ϵ is much larger than the gradient, the regularized normalized fields $\mathbf{n}_\epsilon(u)$ are almost zero and therefore do not have a significant effect of the measures $\mathcal{S}_\times$ or $\mathcal{S}_\circ$, respectively. However, in regions where ϵ is much smaller than the gradients, the regularized normal fields are close to the non-regularized ones (Fig. 2).

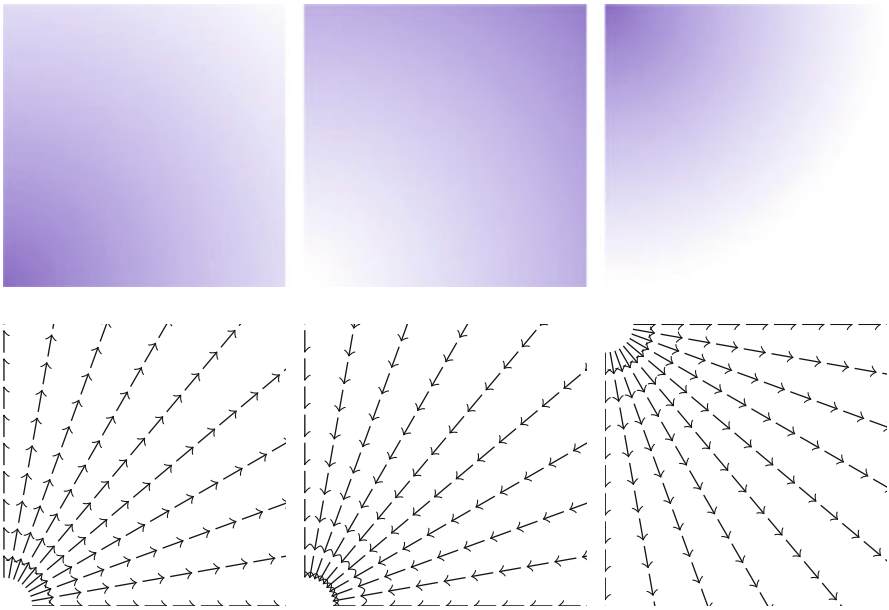


Fig. 2 *Top*: images u_1, u_2, u_3. *Bottom*: $n(u_1), n(u_2), n(u_3)$

Statistical Distance Measures

Several distance measures for pairs of images can be motivated from statistics by considering the images as random variables. In the following, we analyze discrete images from a statistical point of view. For this purpose, we need some elementary statistical definitions. Applications of the following measures are mentioned in section "Morphological Measures":

Correlation Coefficient :

$$\bar{U} := \frac{1}{N^2} \sum_{i,j=1}^{N} U^{ij} \quad \text{and} \quad \mathrm{Var}(U) = \sum_{i,j=1}^{N} (U^{ij} - \bar{U})^2$$

denote the *mean intensity* and *variance* of the discrete image U.

$$\mathrm{Cov}(U_1, U_2) = \sum_{i=1}^{N} \sum_{j=1}^{N} \left(U_1^{ij} - \bar{U}_1\right) \left(U_2^{ij} - \bar{U}_2\right)$$

denotes the *covariance* of two images U_1 and U_2, and the *correlation coefficient* is defined by

$$\rho(U_1, U_2) = \frac{\mathrm{Cov}(U_1, U_2)}{\sqrt{\mathrm{Var}(U_1)\ \mathrm{Var}(U_2)}}.$$

The correlation coefficient is a measure of *linear dependence* of two images. The range of the correlation coefficient is $[-1,\ 1]$, and if $\rho(U_1,\ U_2)$ is close to one, then it indicates that U_1 and U_2 are linearly dependent.

Correlation Ratio: In statistics, the correlation ratio is used to measure the relationship between the statistical dispersion within individual categories and the dispersion across the whole population. The *correlation ratio* is defined by

$$\eta(U_2|U_1) = \frac{\mathrm{Var}(E(U_2|U_1))}{\mathrm{Var}(U_2)},$$

where $E(U_2|U_1)$ is the conditional expectation of U_2 subject to U_1.
To put this into the context of image comparison, let

$$\Omega_t(U_1) := \left\{(i, j)|U_1^{ij} = t\right\}$$

be the discrete level set of intensity $t \in \{0, \ldots, 255\}$. Then the expected value of U_2 on the t-th level set of U_1 is given by

$$E(U_2|U_1 = t) := \frac{1}{\#(\Omega_t(U_1))} \sum_{\Omega_t(U_1)} U_2^{ij},$$

where $\#(\Omega_t(U_1))$ denotes the number of pixels in U_1 with gray value t. Moreover, the according conditional variance is defined by

$$V(U_2|U_1 = t) = \frac{1}{\#(\Omega_t(U_1))} \sum_{\Omega_t(U_1)} \left(U_2^{ij} - E(U_2|U_1 = t)\right)^2.$$

The function

$$\begin{aligned} H(U_1) : \{0, \ldots, 255\} &\to \mathbb{N} \\ t &\mapsto \#(\Omega_t(U_1)) \end{aligned}$$

is called the *discrete histogram* of U_1.
The correlation ratio is nonsymmetric, that is, $\eta(Y|X) \neq \eta(X|Y)$, and takes values in $[0, 1]$. It is a measure of *(non)linear dependence* between two images. If $U_1 = U_2$, then the correlation ratio is maximal.

Variance of Intensity Ratio, Ratio Image Uniformity: This measure is based on the definition of similarity that two images are similar, if the factor $R^{ij}\ (U_1,\ U_2) = U_1^{ij}/U_2^{ij}$ has a small variance. The *ratio image uniformity* (or normalized variance of the intensity ratio) can be calculated by

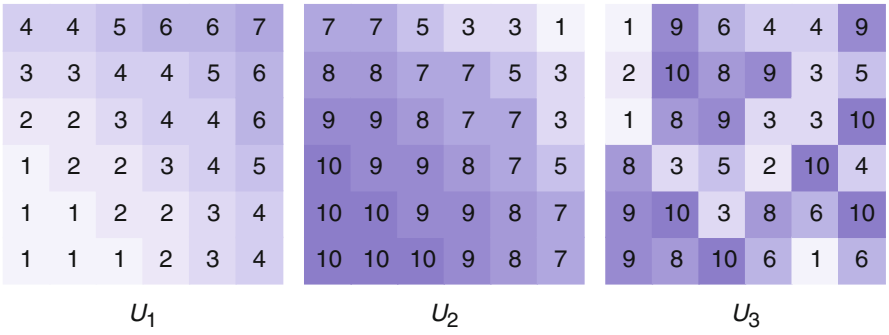


Fig. 3 Images for Examples 2 and 6. Note that there is a dependence between U_1 and U_2: $U_2 \sim 11 - (U_1)^3$

Table 1 Comparison of the different pixel-based similarity measures. The images U_1, U_2 are related in a nonlinear way; this is reflected in a correlation ratio of 1. We see that the variance of intensity ratio is not symmetric and not significant to make a statement on a correlation between the images

	U_1, U_2	U_2, U_1	U_2, U_3	U_3, U_2	U_3, U_1	U_1, U_3
Correlation coefficient	**−0.98**	**−0.98**	0.10	0.10	−0.14	−0.14
Correlation ratio	**1.00**	**1.00**	0.28	0.32	0.29	0.64
Variance of intensity ratio	1.91	2.87	2.25	1.92	**3.06**	0.83

$$\mathrm{RIU}(U_1, U_2) = \frac{\mathrm{Var}(R)}{\bar{R}}.$$

It is not symmetric.

Example 2. Consider the discrete images U_1, U_2, and U_3 in Fig. 3. Table 1 shows a comparison of the different similarity measures. The variance of the intensity ratio is insignificant and therefore cannot be used to determine similarities. The correlation ratio is maximal for the pairing U_1, U_2, and in fact there is a functional dependence of the intensity values of U_1 and U_2. However, the dependence of the intensity values of U_1 and U_2 is nonlinear; hence, the absolute value of the correlation coefficient (measure of linear dependence) is close to one, but not identical to one.

Statistical Distance Measures (Density Based)

In general, two images of the same object but of different modalities have a large L^p, l^p distance. Hence, the idea is to apply statistical tools that consider images as similar if there is some statistical dependence. Statistical similarity measures are able to compare probability density functions. Hence, we first need to relate images

to density functions. Therefore, we consider an image as a random variable. The basic terminology of random variables is as follows:

Definition 1. *A continuous random variable* is a real-valued function $X : \Omega^S \to \mathbb{R}$ defined on the *sample space* Ω^S. For a *sample* x, $X(x)$ is called *observation*.

Remark 1 (Images as Random Variables). When we consider an image $u : \Omega \to \mathbb{R}$ as a continuous random variable, the sample space is Ω. For a sample $x \in \Omega$, the observation $u(x)$ is the intensity of u at x.

Regarding the intensity values of an image as an observation of a random process allows us to compare images via their intrinsic probability densities. Since the density *cannot* be calculated directly, it has to be estimated. This is outlined in section "Density Estimation". There exists a variety of distance measures for probability densities (see, for instance, [31]). In particular, we review f-divergences in section "Csiszár-Divergences (f-Divergences)" and explain how to use the f-information as an image similarity measure in section "f-Information".

Density Estimation

This section reviews the problem of *density estimation*, which is the construction of an estimate of the density function from the observed data.

Definition 2. Let $X : \Omega^S \to \mathbb{R}$ be a random variable, that is, a function mapping the (measurable) sample space Ω^S of a random process to the real numbers.

The *cumulated probability density function* of X is defined by

$$P(t) := \frac{1}{\text{meas}(\Omega^S)} \text{meas}\{x : X(x) \leq t\}\ t \in \mathbb{R}.$$

The *probability density* function p is the derivative of P.

The *joint cumulated probability density function* of two random variables X_1, X_2 is defined by

$$\hat{P}(t_1, t_2) := \frac{1}{\text{meas}(\Omega^S)^2} \text{meas}\{(x_1, x_2) : X_1(x_1) \leq t_1, X_2(x_2) \leq t_2\}\ t_1, t_2\ \in \mathbb{R}.$$

The *joint probability density* function $\hat{p}$ satisfies

$$\hat{P}(t_1, t_2) = \int_0^{t_1} \int_0^{t_2} \hat{p}(s_1, s_2) ds_1\ ds_2.$$

Remark 2. When we consider an image $u : \Omega \to \mathbb{R}$ a random variable with sample space Ω, we write $p(u)(t)$ for the probability density function of the image u. For the joint probability of two images u_1 and u_2, we write $\hat{p}(u_1, u_2)(t_1, t_2)$ to emphasize, as above, that the images are considered as random variables.

The terminology of Definition 2 is clarified by the following one-dimensional example:

Example 3. Let $\Omega := [0,\ 1]$ and

$$\begin{aligned} u : \Omega &\to [0, 255] \\ x &\to 255x^2. \end{aligned}$$

The cumulated probability density function

$$P :\ [0,\ 255] \to [0,\ 1]$$

is obtained by integration:

$$P(t) := \text{meas}\{x : 255\ x^2 \le t\} = \text{meas}\left\{x : x \le \sqrt{\frac{t}{255}}\right\} = \int_0^{\sqrt{\frac{t}{255}}} 1\ dx = \sqrt{\frac{t}{255}}.$$

The probability density function of u is given by the derivative of P, which is

$$p(u)(t) = \frac{1}{2\sqrt{255}} \frac{1}{\sqrt{t}}.$$

In image processing, it is common to view the discrete image U (or u_N as in (1)) as an approximation of an image u. We aim for the probability density function of u, which is approximated via kernel density estimation using the available information of u, which is U. A kernel histogram is the normalized probability density function according to the discretized image U, where for each pixel a *kernel function* (see (3)) is superimposed. Kernel functions depend on a parameter, which can be used to control the smoothness of the kernel histogram.

We first give a general definition of kernel density estimation:

Definition 3 (Kernel Density Estimation). Let $t_1,\ t_2,\ \ldots,\ t_M$ be a sample of M independent observations from a measurable real random variable X with probability density function p. A *kernel density approximation* at t is given by

$$p_\sigma(t) = \frac{1}{M} \sum_{i=1}^{M} k_\sigma(t - t_i), \quad t \in [0, 255]$$

where k_σ is a kernel function with bandwidth σ. p_σ is called *kernel density approximation* with parameter σ.

Let $t_1, t_2, \ldots, t_M$ and $s_1, s_2, \ldots, s_M$ be samples of M independent observations from measurable real random variables X_1, X_2 with joint probability density function $\hat{p}$; then a *joint kernel density approximation* of $\hat{p}$ is given by

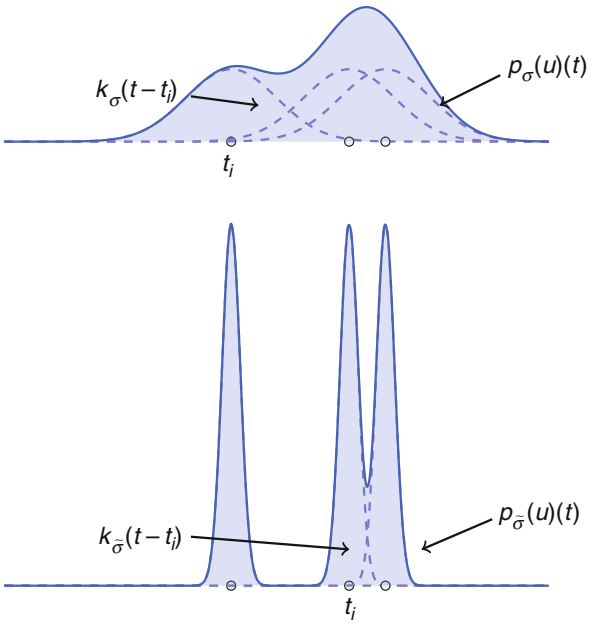


Fig. 4 Density estimate for different parameters σ

$$\hat{p}_\sigma(s,t) = \frac{1}{M}\sum\nolimits_{i=1}^{M} K_\sigma(s-s_i, t-t_i),$$

where $K_\sigma(s,\ t)$ is a two-dimensional kernel function.

Remark 3 (Kernel Density Estimation of an Image, Fig. 4). Let u be a continuous image, which is identified with a random variable. Moreover, let U be $N \times N$ samples of u. In analogy to Definition 3, we denote the *kernel density estimation* based on the discrete image U, by

$$p_\sigma(t) = \frac{1}{N^2}\sum\nolimits_{i,j=1}^{N} k_\sigma(t-U^{ij})$$

and remark that for u_N as in (1)

$$p_\sigma(u_N)(t) := \int_\Omega k_\sigma(t-u_N(x))dx = \frac{1}{N^2}\sum\nolimits_{i,j=1}^{N} k_\sigma(t-U^{ij}). \tag{3}$$

The joint kernel density of two images u_1, u_2 with observations U_1 and U_2 is given by

$$\hat{p}_\sigma(s,t) = \frac{1}{N^2}\sum\nolimits_{i,j=1}^{N} K_\sigma\left(s-U_1^{ij}, t-U_2^{ij}\right),$$

where $K_\sigma(s,\ t) = k_\sigma(s)k_\sigma(t)$ is the two-dimensional kernel function. Moreover, we remark that for $u_{1,N}$, $u_{2,N}$

$$\begin{aligned}\hat{p}_\sigma(u_{1,N}, u_{2,N})(s,t) &:= \int_\Omega K_\sigma(s - u_{1,N}(x), t - u_{2,N}(x))\ dx \\ &= \frac{1}{N^2}\sum\nolimits_{i,j=1}^{N} K_\sigma\left(s - U_1^{ij}, t - U_2^{ij}\right).\end{aligned}$$

In the following, we review particular kernel functions and show that standard histograms are kernel density estimations.

Example 4. Assume that $u_i : \Omega \to [0,\ 255]$, $i = 1,\ 2$ are continuous images, with discrete approximations $u_{i,N}$ as in (1):

- We use the joint density kernel $K_\sigma(s,\ t) := k_\sigma(s)k_\sigma(t)$, where k_σ is the **normalized Gaussian kernel** of variance σ. Then for $i = 1,\ 2$, the estimates for the marginal densities are given by

$$p_\sigma(u_{i,N})(t) := \int_\Omega k_\sigma(u_{i,N}(x) - t)\ dx = \frac{1}{\sqrt{2\pi}\sigma}\int_\Omega \exp\left(\frac{-(u_{i,N}(x) - t)^2}{2\sigma^2}\right) dx,$$

  and the joint density approximation reads as follows:

$$\begin{aligned}\hat{p}_\sigma(s,t) &:= \int_\Omega K_\sigma((u_1(x), u_2(x)) - (s,t))\ dx \\ &= \tfrac{1}{2\pi\sigma^2}\int_\Omega \exp\left(\tfrac{-(u_{1,N}(x)-s)^2}{2\sigma^2}\right)\exp\left(\tfrac{-(u_{2,N}(x)-t)^2}{2\sigma^2}\right) dx\end{aligned}$$

- **Histograms:** Assume that U only takes values in $0, 1, \ldots, 255$. When we choose the characteristic function $\chi_{[-\sigma,\sigma)}$, with $\sigma = \frac{1}{2}$ as kernel function, we obtain the density estimate

$$\begin{aligned}p_{\chi,\sigma}(t) &= \int_\Omega \chi_{[-\sigma,\sigma]}(u(x) - t)\ dx \\ &= \text{meas}\{x : t - \sigma \le u(x) < t + \sigma\} \\ &= \text{size of pixel} \times \text{number of pixels with value } \lfloor t + \sigma \rfloor\end{aligned}$$

  Hence, $p_{\chi,\sigma}$ corresponds with the histogram of the discrete image.

Example 5. We return to Example 3. The domain $\Omega = [0,\ 1]$ is partitioned into N equidistant pieces. Let

$$u_N := \sum\nolimits_{i=1}^{N}\left(\int_{\frac{i-1}{N}}^{\frac{i}{N}} u(x)dx\right)\chi_{[\frac{i-1}{N},\frac{i}{N}]}.$$

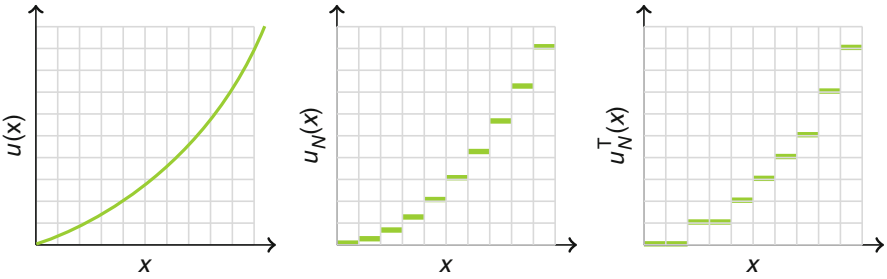


Fig. 5 Original u and discretized versions u_N and ${u_N}^T$

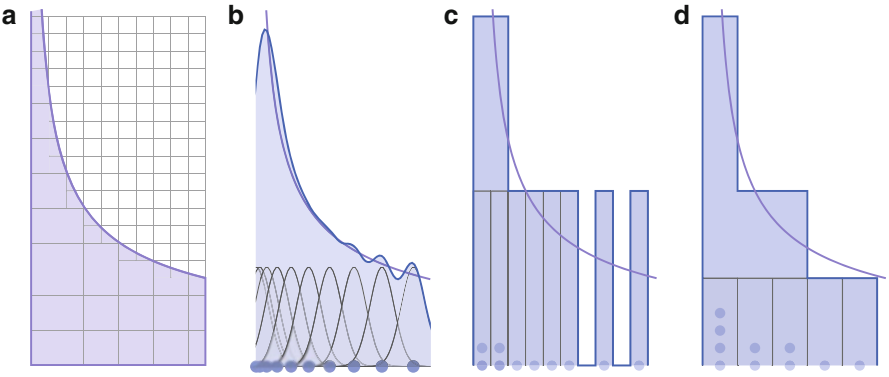


Fig. 6 (**a**) Density from the original image u. (**b**) Density estimation with Gaussian-kernel based on $u_N (N = 10)$, with $\sigma = 0.07$, (**c**, **d**) normalized histogram, based on ${u_N}^T$, with $\sigma = 0.05,\ 0.1$

Moreover, we consider the piecewise function ${u_N}^T$ represented in Fig. 5. The density according to u, denoted by $p(u)$, and the kernel density estimates of u_N and ${u_N}^T$ are represented in Fig. 6. They resemble the actual density very well.

Csiszár Divergences (*f*-Divergences)

The concept of f-divergences has been introduced by Csiszár in [5] as a generalization of Kullback's I-divergence and Rényi's I-divergence and at the same time by Ali and Silvey [1]. In probability calculus, f-divergences are used to measure the distances between probability densities.

Definition 4. Set $\mathcal{F}_0 := \{f : [0,\ \infty) \to \mathbb{R} \cup \{\infty\} \ : f$ is convex in $[0,\ \infty)$, continuous at 0, and satisfies $f(1) = 0\}$ and

$$V_{\text{pdf}} := \left\{ p \in L^1(\mathbb{R}) : p \geq 0, \int_{\mathbb{R}} p(t)\ dt = 1 \right\}.$$

Let $g_1, g_2 \in V_{\text{pdf}}$ be probability density functions. The f-divergence between g_1, g_2 is given by

$$\begin{aligned} \mathcal{D}_f : V_{\text{pdf}} \times V_{\text{pdf}} &\to [0, \infty) \\ (g_1, g_2) &\to \int_{\mathbb{R}} g_2(t) f\left(\tfrac{g_1(t)}{g_2(t)}\right) dt. \end{aligned} \tag{4}$$

Remark 4.

- In (4), the integrand at positions t where $g_2(t) = 0$ is understood in the following sense:

$$0 f\left(\frac{g_1(t)}{0}\right) := \lim_{\bar{t} \searrow 0} \left(\bar{t} f\left(\frac{g_1(t)}{\bar{t}}\right)\right), \quad t \in \mathbb{R}.$$

- In general, f-divergences are not symmetric, unless there exists some number c such that the generating f satisfies $f(x) = x\, f\left(\frac{1}{x}\right) + c(x-1)$.

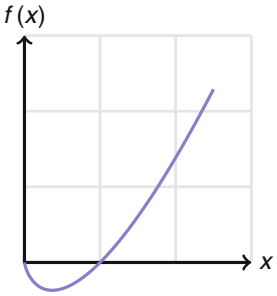
Examples for f-Divergences: We list several f-divergences that have been used in literature (see [6, 12] and references therein).

The Kullback-Leibler divergence is the f-divergence with $f(x) = x \log(x)$

$$\mathcal{D}_f(g_1, g_2) = \int_{\mathbb{R}} g_1(t) \log\left(\frac{g_1(t)}{g_2(t)}\right) dt.$$

Jensen-Shannon divergence is the symmetric Kullback-Leibler divergence:

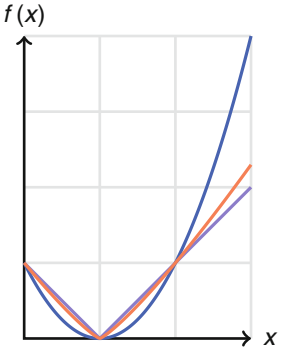
$$\mathcal{D}_f(g_1, g_2) = \int_{\mathbb{R}} \left(g_1(t) \log\left(\frac{g_1(t)}{g_2(t)}\right) + g_2(t) \log\left(\frac{g_2(t)}{g_1(t)}\right)\right) dt.$$



χ^s-Divergences: These divergences are generated by

$$f^s(x) = |x-1|^s, \quad s \in [1, \infty)$$

and have the form



$$\mathcal{D}_f(g_1, g_2) = \int_{\mathbb{R}} g_2(t) \left| \frac{g_1(t)}{g_2(t)} - 1 \right|^s = \int g_2^{1-s}(t) |g_1(t) - g_2(t)|^s dt.$$

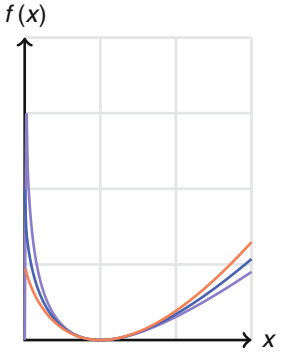
The χ^1-divergence is a metric. The most widely used out of this family of χ^s divergences is the χ^2-divergence (Pearson).

Dichotomy Class Divergences: The generating function of this class is given by

$$f(x) = \begin{cases} x - 1 - \ln(x) & \text{for } s = 0, \\ \frac{1}{s(1-s)}(sx + 1 - s - x^s) & \text{for } s \in \mathbb{R}\backslash\{0, 1\}, \\ 1 - x + x \ \ln(x) & \text{for } s = 1. \end{cases}$$

The parameter $s = \frac{1}{2}$ provides a distance, namely, the Hellinger metric

$$\mathcal{D}_f(g_1, g_2) = 2 \int_{\mathbb{R}} \left(\sqrt{g_1(t)} - \sqrt{g_2(t)} \right)^2 dt.$$



Matsushita's Divergences: The elements of this class, which is generated by the function

$$f(x) = |1 - x^s|^{\frac{1}{s}}, \ 0 < s \leq 1,$$

are prototypes of metric divergences. The distance is given by

$$d(g_1, g_2) = (\mathcal{D}_f(g_1, g_2))^s$$

where

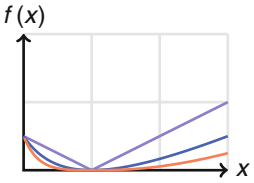
$$\mathcal{D}_f(g_1, g_2) = \int_{\mathbb{R}} g_1(t) \left| 1 - \left(\frac{g_2(t)}{g_1(t)} \right)^s \right|^{\frac{1}{s}} dt.$$

Puri-Vincze Divergences: This class is generated by the functions

$$f(x) = \frac{|1 - x|^s}{2(x+1)^{s-1}}, \quad s \in [1, \infty).$$

For $s = 2$, we obtain the triangular divergence

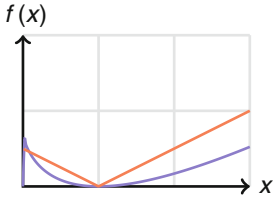
$$\mathcal{D}_f(g_1, g_2) = \int_{\mathbb{R}} \frac{(g_2(t) - g_1(t))^2}{g_2(t) + g_1(t)} dt.$$



Divergences of Arimoto Type: Generated by the functions

$$f(x) = \begin{cases} \frac{s}{s-1} \left((1 + x^s)^{\frac{1}{s}} - 2^{\frac{1}{s}-1}(1 + x) \right) & \text{for } s \in (0, \infty) \backslash \{1\} \\ (1 + x)\ln(2) + x \ln(x) - (1 + x)\ln(1 + x) & \text{for } s = 1 \\ \frac{1}{2}|1 - x| & \text{for } s = \infty. \end{cases}$$

For $s = \infty$, the divergence is proportional to the χ^1 divergence. For $s \in (0, \infty) \backslash \{1\}$, we obtain



$$\mathcal{D}_f(g_1, g_2) = \int_{\mathbb{R}} \frac{s}{s-1} \left(\sqrt[s]{g_1^s(t) + g_2^s(t)} - 2^{\frac{1-s}{s}} (g_1(t) + g_2(t)) \right) dt.$$

Moreover, this class provides the distances

$$d(g_1, g_2) = (\mathcal{D}_f(g_1, g_2))^{\min\{s, \frac{1}{s}\}} \quad \text{for } s \in (0, \infty).$$

f-Information

In the following, we review the f-information for measuring the distance between probability densities. The most important f-information measure is the *mutual information*.

The notion of *information gain* induced by simultaneously observing two probability measures compared to their separate observations is tightly related to divergences. It results from quantifying the information content of the joint measure in comparison with the product measure.

This motivation leads to the following definition.

Definition 5 (f-Information for Images). For $f \in \mathcal{F}_0$ (see Definition 4), we define the **f-information** of $u_1, u_2 \in L^\infty(\Omega)$ by

$$I_f(u_1, u_2) := \mathcal{D}_f(p(u_1)p(u_2), p(u_1, u_2)),$$

where the $p(u_i)$ is the probability density of u_i, as introduced in section "Density Estimation."

Additionally, we define the **f-entropy** of an image u_1 by

$$H_f(u_1) := I_f(u_1, u_1).$$

In analogy to independent probability densities, we call two images u_1, u_2 **independent** if there is no information gain, that is,

$$p(u_1, u_2) = p(u_1)p(u_2).$$

Remark 5. The f-information has the following properties:

- **Symmetry:** $I_f(u_1,\ u_2) = I_f(u_2,\ u_1)$.
- **Bounds:** $0 \leq I_f(u_1, u_2) \leq \min\{H_f(u_1),\ H_f(u_2)\}$.
- $I_f(u_1, u_2) = 0$ if and only if u_1, u_2 are mutually independent.

The definition of f-information does not make assumptions on the relationship between the image intensities (see [38] for discussion). It does neither assume a linear nor a functional correlation but only a predictable relationship. For more information on f-information, see [36].

Example 6. The most famous examples of f-informations are the following:

Mutual/Shannon Information: For $f(x) = x \ln x$, we obtain

$$I_f(u_1, u_2) = \int_{\mathbb{R}} \int_{\mathbb{R}} p(u_1, u_2)(t_1, t_2) \ln \left(\frac{p(u_1, u_2)(t_1, t_2)}{p(u_1)(t_1)\, p(u_2)(t_2)} \right) \, dt_1 \, dt_2,$$

with Shannon entropy

$$H_f(u) = \int_{\mathbb{R}} p(u)(t) \ln \left(\frac{1}{p(u)(t)} \right) \, dt,$$

joint entropy

$$H_f(u_1, u_2) = -\int_{\mathbb{R}} \int_{\mathbb{R}} p(u_1, u_2)(t_1, t_2) \ln \left(p(u_1, u_2)(t_1, t_2) \right) \, dt_1 \, dt_2,$$

conditional entropy

$$H_f(u_2|u_1) = \int_{\mathbb{R}} p(u_1)(t) H_f(u_2|u_1 = t) \, dt,$$

and relative entropy (the Kullback-Leibler divergence)

$$H_f(u_1|u_2) = \int_{\mathbb{R}} p(u_1)(t) \ln \left(\frac{p(u_1)}{p(u_2)} \right).$$

The relative entropy is not symmetric. Maes et al. [25] and Studholme et al. [35] both suggested the use of joint entropy for multimodal image registration. Maes et al. demonstrate the robustness of registration, using mutual information with respect to partial overlap and image degradation, such as noise and intensity inhomogeneities.

Hellinger Information: For $f(x) = 2x - 2 - 4\sqrt{x}$ (see also Dichotomy Class in section "Csiszár-Divergences (f-Divergences)"), we obtain

$$I_f(u_1,u_2) = \int_{\mathbb{R}}\int_{\mathbb{R}} \left(\sqrt{p(u_1,u_2)(t_1,t_2)} - \sqrt{p(u_1)(t_1)\,p(u_2)(t_2)}\right)^2 dt_1\, dt_2,$$

with Hellinger entropy

$$H_f(u) = 2\left(1 - \int_{\mathbb{R}} (p(u)(t))^{\frac{3}{2}} dt\right).$$

Both are bounded from above by 2.

For measuring the distance between discrete images U_1 and U_2, it is common to map the images via kernel estimation to continuous estimates of their intensity value densities $p_\sigma(u_{i,N})$, where $p_\sigma(u_{i,N})$ is as defined in (3). The difference between images is then measured via the distance between the associated estimated probability densities.

Example 7. For U_i, $i = 1,2,3$ as in Fig. 3, let $u_{i,N}$ be the corresponding piecewise constant functions. Note that U_1 and U_2 are somehow related. In other words, they are highly dependent on each other, so we can expect a low information value. Comparing the images point-wise with least squares shows a higher similarity value for U_2 and U_3 than for U_1 and U_2.

For the ease of presentation, we work with histograms. Recall that the estimated probability function $p_\sigma(u_{i,N})$ is equal to the normalized histogram of U_i. The histograms (connected to the marginal density densities) are given by

	1	2	3	4	5	6	7	8	9	10
$H(U_1)$	6	7	6	9	3	4	1	0	0	0
$H(U_2)$	1	0	4	0	3	0	9	6	7	6
$H(U_3)$	3	2	5	3	2	4	0	5	6	6

In order to calculate the information measures, we calculate the joint histograms of U_1, U_2, U_3, that is, $JH(U_1, U_2) : (s, t) \to$ number of pixels such that $U_1^{ij} = s$ and $U_2^{ij} = t$ (see Tables 2 and 3).

The entries in the joint histogram of U_1, U_2 are located along a diagonal, whereas the entries of the other two joint histograms are spread all over. Hence, we can observe the dependence of $p_\sigma(u_{1,N})$, $p_\sigma(u_{2,N})$ already by looking at the joint histogram. Next, we calculate the Hellinger and the mutual information. For the f-entropies, we obtain

	U_1	U_2	U_3
Mutual entropy	1.91	1.91	2.13
Hellinger entropy	1.17	1.17	1.30

Table 2 Joint histograms of U_2, U_3 and U_3, U_1. The entries are disperse; this will be reflected in a lower f-information as in the case for U_1, U_2

$JH(U_1,U_2)$	1	2	3	4	5	6	7	8	9	10	$H(U_2)$
1							1				1
2											0
3						4					4
4											0
5					3						3
6											0
7				9							9
8			6								6
9		7									7
10	6										6
$H(U_1)$	6	7	6	9	3	4	1	0	0	0	

Table 3 Joint histograms of U_2, U_3 and U_3, U_1. The entries are disperse; this will be reflected in a lower f-information as in the case for U_1, U_2

$JH(U_2,U_3)$	1	2	3	4	5	6	7	8	9	10	$H(U_3)$
1							1	1	1		3
2								2			2
3					1		2		2		5
4			2		1						3
5			1						1		2
6					1		1	1	1		4
7											0
8							1		2	2	5
9	1						2	1		2	6
10			1				2	1		2	6
$H(U_2)$	1	0	4	0	3	0	9	6	7	6	

$JH(U_3,U_1)$	1	2	3	4	5	6	7	8	9	10	$H(U_1)$
1								2	2	2	6
2	1		2		1	1		2			7
3	1	2				1			1	1	6
4	1		2			1		1	2	2	9
5			1	1		1					3
6				2	1					1	4
7									1		1
8											0
9											0
10											0
$H(U_3)$	3	2	5	3	2	4	0	5	6	6	

and for the f-information measures:

	(U_1, U_2)	(U_2, U_3)	(U_3, U_1)
Mutual information	1.91	0.74	0.74
Hellinger information	1.17	0.57	0.57
Sum of least squares	31.44	14.56	21.56

Indeed, in both cases (Hellinger and mutual information), U_1, U_2 (high f-information value) can be considered as more similar than U_1 and U_3, whereas the least squares value between U_1, U_2 is the highest, meaning that they differ at most.

We can observe in Example 7 that the values of f-information differ a lot by different choices of the function f. Moreover, it is not easy to interpret the values; hence, one is interested in calculating normalized values:

Normalized Mutual Information: Studholme [35] proposed a normalized measure of mutual information. Normalized f-information is defined by

$$\mathrm{NI}_f(u_1, u_2) := \frac{H_f(u_1) + H_f(u_2)}{I_f(u_1, u_2)}.$$

If $u_1 = u_2$, then the normalized f-information is minimal with value 2.

Entropy Correlation Coefficient: Collignon and Maes [25] suggested the use of the entropy correlation coefficient, another form of normalized f-information (Table 4):

$$H_f\mathrm{CC}(u_1, u_2) = \frac{2I_f(u_1, u_2)}{H_f(u_1) + H_f(u_2)} = 2 - \frac{2}{\mathrm{NI}_f(u_1, u_2)}.$$

The entropy correlation coefficient is one if $u_1 = u_2$ and zero if u_1 and u_2 are completely independent.

Exclusive f-Information: It is defined by

$$EI_f(u_1, u_2) := H_f(u_1) + H_f(u_2) - 2I_f(u_1, u_2).$$

Note that the exclusive f-information is **minimal** for $u_1 = u_2$.

Distance Measures Including Statistical Prior Information

Most multimodal measures used in literature do not consider the underlying image context or other statistical prior information on the image modalities. Recently, several groups developed similarity measures that incorporate such information:

- Leventon and Grimson [23] proposed to learn *prior information* from training data (registered multimodal images) by estimating the joint intensity distributions

Table 4 Comparison of measures composed by f-information and f-entropies

Mutual information	(u_1, u_2)	(u_2, u_3)	(u_3, u_1)
Normalized[a]	**2.00**	5.32	5.32
Entropy correlation coefficient	**1.00**	0.38	0.38
Exclusive[a]	**0.00**	2.46	2.46
Hellinger information	(u_1, u_2)	(u_2, u_3)	(u_3, u_1)
Normalized[a]	**2.00**	4.36	4.36
Entropy correlation coefficient	**1.00**	0.46	0.46
Exclusive[a]	**0.00**	1.34	1.34

[a]Normalized and exclusive informations are minimal if the images are equal, whereas the entropy correlation coefficient is maximal

of the training images. Based on this paper, Chung et al. [4] proposed to use the Kullback-Leibler distance to compare the learned joint intensity distribution, with the joint intensity distribution of the images, in order to compare multimodal images. This idea was extended by Guetter et al. [14], who combine mutual information with the incorporation of learned prior knowledge with a Kullback-Leibler term.

As a follow-up of their ideas, we suggest the following type of generalized similarity measures: Let p_σ^l be the learned joint intensity density (learned from the training data set) and $\alpha \in [0,\ 1]$. For $f \in \mathcal{F}_0$, define

$$\mathcal{S}_{\alpha,\sigma}(u_1,u_2) := \alpha \mathcal{D}_f(p_\sigma^l,\, p_\sigma(u_1,u_2)) + (1-\alpha)\underbrace{\mathcal{D}_f(p_\sigma(u_1,u_2),\, p_\sigma(u_1)p_\sigma(u_2))}_{I_f(u_1,u_2)}.$$

- Instead of using a universal but a priori fixed similarity measure, one can *learn a similarity* measure in a discriminative manner. The methodology proposed by Lee et al. [22] uses a learning algorithm that constructs a similarity measure based on a set of preregistered images.

3 Mathematical Models for Variational Imaging

In the following, we proceed with an abstract setting. We are given a physical model F, which in mathematical terms is an operator between spaces U and V. For given data $v \in V$, we aim for solving the operator equation

$$F(\Phi) = v.$$

In general, the solution is not unique, and we aim for finding the solution with *minimal energy*, that is, we aim for a minimizer of the constraint optimization problem

$$\mathcal{R}(\Phi) \to \min \text{ subject to } F(\Phi) = v.$$

In practice, a complication of this problem is that only approximate (noisy) data $v^\delta \in V$ of v is available. To take into account uncertainty of the data, it is then intuitive to consider the following constrained optimization problem instead:

$$\mathcal{R}(\Phi) \to \min \text{ subject to } \left\| F(\Phi) - v^\delta \right\|^2 \le \delta, \tag{5}$$

where δ is an upper bound for the approximation error $v - v^\delta$. It is known that solving (5) is equivalent to minimizing the Tikhonov functional,

$$\Phi \to \frac{1}{2}\left\| F(\Phi) - v^\delta \right\|^2 + \alpha \mathcal{R}(\Phi), \tag{6}$$

where $\alpha > 0$ is chosen according to Morozov's discrepancy principle [19].

For the formulation of the constrained optimization problem, the Tikhonov method, respectively, it is essential that $F(\Phi)$ and v, v^δ, respectively, represent data of the same modality. If $F(\Phi)$ and the data, which we denote now by w, are of different kind, then it is intuitive to use a multimodal similarity measure $\mathcal{S}^r$, instead of the least squares distance, which allows for comparison of $F(\Phi)$ and w. Consequently, we consider the multimodal variational method, which consists in minimization of

$$\Phi \to \mathcal{T}_{\alpha,w^\delta}(\Phi) := \mathcal{S}^r(F(\Phi), w^\delta) + \alpha\mathcal{R}(\Phi), \quad \alpha > 0.$$

In the limiting case, that is, for $\delta \to 0$, one aims for recovering an $\mathcal{RS}^r$-minimizing solution $\Phi^\dagger$ if

$$\mathcal{R}(\Phi^\dagger) = \min\{\mathcal{R}(\Phi) : \Phi \in \mathcal{A}\} \text{ where } \mathcal{A} = \{\Phi : \Phi = \arg\min\{\mathcal{S}^r(F(\cdot), w)\}\}.$$

To take into account priors in the Tikhonov regularization, the standard way is again by a least squares approach. In this case, for regularization the least squares functional

$$\Phi \to \mathcal{R}_1(\Phi) = \frac{1}{2}\|\Phi - \Phi_0\|^2$$

is added to $\frac{1}{2}\left\|F(\Phi) - v^\delta\right\|^2$ (see, e.g., [8]). In analogy, we consider the regularization functional (6) and incorporate priors by adding generalization of the functional $\mathcal{R}_1(\Phi)$. Taking into account prior information Ψ_0, which might come, for instance, from another modality, this leads to the following class of generalized Tikhonov functionals:

$$\mathcal{T}_{\alpha,\beta}^{w^\delta,\Psi_0}(\Phi) := \mathcal{S}^r(F(\Phi), w^\delta) + \alpha\mathcal{R}(\Phi) + \beta\mathcal{S}^p(\Phi, \Psi_0).$$

Here $\mathcal{S}^p$ is an appropriate multimodal similarity measure. In the limiting case, that is, for $\delta \to 0$, one aims for recovering an $\gamma - \mathcal{RS}^r\mathcal{S}^p$-minimizing solution $\Phi^\dagger$ if

$$\begin{aligned}\mathcal{R}(\Phi^\dagger) + \gamma\mathcal{S}^p(\Phi^\dagger, \Psi_0) = \min\{\mathcal{R}(\Phi) + \gamma\mathcal{S}^p(\Phi^\dagger, \Psi_0) : \Phi \in \mathcal{A}\} \text{ where}\\ \mathcal{A} = \{\Phi : \Phi = \arg\min\{\mathcal{S}^r(F(\cdot), w)\}\}.\end{aligned}$$

The γ-parameter balances between the amount of prior information and regularization and satisfies $\gamma = \lim_{\alpha,\beta\to 0}\frac{\beta}{\alpha}$. For theoretical results on existence of minimizing elements of the functionals and convergence, we refer to [11, 30].

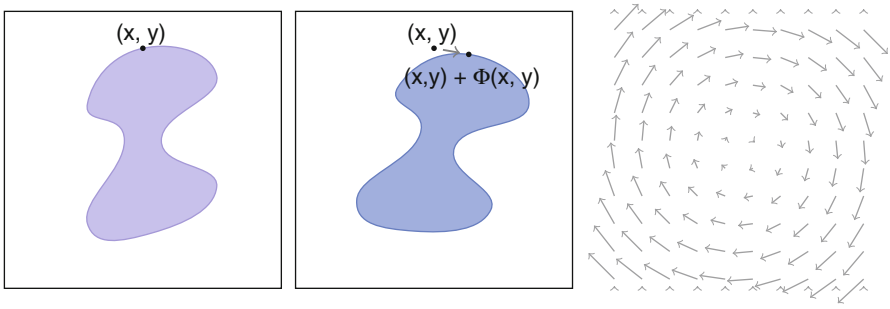


Fig. 7 *Left*: images u_R, u_T, *right*: deformation field Φ

4 Registration

In this section, we review variational methods for *image registration*. This problem consists in determining a spatial transformation (vector field) Φ that aligns pixels of two images u_R and u_T in an optimal way (Fig. 7). We use the terminology *reference image* for u_R and *template image* for u_T, where both are assumed to be compactly supported functions in Ω. That is, we consider the problem of determining the optimal transformation, which minimizes the functional

$$u \to \mathcal{S}^r(u_T \circ (id + \Phi), u_R). \tag{7}$$

To establish the context to the inverse problem setting, we use the setting $F(\Phi) = u_T\ (id + \Phi)$ and $w^\delta = u_R$. In general, the problem of minimizing (7) is ill posed. The Tikhonov-type variational regularization for registration then consists in minimization of the functional

$$\Phi \to \mathcal{S}^r(u_T \circ (id + \Phi), u_R) + \alpha \mathcal{R}(\Phi) \tag{8}$$

(we do not consider constrained registration here, but concentrate on the Tikhonov regularization).

Image registration (also of voxel (3D) data) is widely used in medical imaging, for instance, for monitoring and evaluating tumor growth, disease development, and therapy evaluation.

Variational methods for registration differ by the choice of the regularization functional $\mathcal{R}$ and the similarity measure $\mathcal{S}^r$. There exists a variety of similarity measures that are used in practice. For some surveys, we refer to [17, 27, 32].

The regularization functional $\mathcal{R}$ typically involves differential operators. In particular, for nonrigid registration, energy functionals from elasticity theory and fluid dynamics are used for regularization.

The optimality condition for a minimizer of (8) reads as follows:

$$\alpha D_\Phi(\mathcal{R}(\Phi), \Psi) + D_\Phi(\mathcal{S}^r(u_T \circ (id + \Phi), u_R), \Psi) = 0 \quad \text{for all } \Psi \in U, \tag{9}$$

where $D_\Phi(\mathcal{T}, \Psi)$ denotes the directional derivative of a functional $\mathcal{T}$ in direction Ψ. The left-hand side of the equation is the steepest descent functional of the energy functional (8). In the following, we highlight some steepest descent functionals according to variational registration methods.

Example 8 (Elastic Registration with L^2-Norm-Based Distance Function). Set $\alpha = 1$, $\mathcal{S}^r(v_1, v_2) = \frac{1}{2}\|v_1 - v_2\|_{L^2}^2$. We consider an elastic regularization functional of the form

$$\mathcal{R}(\Phi) = \int_\Omega \sum_{i=1}^2 \sum_{j=1}^2 \left(\frac{\lambda}{2} \frac{\partial}{\partial x_i} \Phi^i \frac{\partial}{\partial x_j} \Phi^j + \frac{\mu}{4} \left(\frac{\partial}{\partial x_j} \Phi^i + \frac{\partial}{\partial x_i} \Phi^j \right)^2 \right),$$

where $\lambda, \mu \geq 0$ are Lamé parameters and $\Phi = (\Phi^1, \Phi^2)$. λ is adjusted to control the rate of growth or shrinkage of local regions within the deforming template, and μ is adjusted to control shearing between adjacent regions of the template [3]. In this case, the optimality condition for minimizing $\alpha\mathcal{R}(\Phi) + \mathcal{S}^r(u_T \circ (id + \Phi))$, given by (9), is satisfied if Φ solves the following PDE:

$$\mu \Delta \Phi(x) + (\mu + \lambda) \nabla(\nabla \cdot \Phi(x)) = \underbrace{-\frac{1}{\alpha}(u_T(x + \Phi(x)) - u_R(x)) \nabla u_T(x + \Phi(x))}_{\frac{\partial}{\partial \Phi}\mathcal{S}^r}.$$

Here $\Delta\Phi = (\Delta\Phi^1, \Delta\Phi^2)$ and $D_\Phi(\mathcal{S}^r(u_T \circ (id + \Phi), u_R), \Psi) = \int_\Omega \frac{\partial}{\partial \Phi}\mathcal{S}^r \cdot \Psi$. This partial differential equation is known as linear elastic equation and is derived assuming small angles of rotation and small linear deformations. When large displacements are inherent, it is not applicable [2, 13, 18, 24, 29, 40].

Example 9 (Elastic Registration with f-Information). Assume that $k_\sigma \in \mathcal{C}^1(\mathbb{R}, \mathbb{R})$ is some kernel density function. Moreover, let $K_\sigma(s, t) = k_\sigma(s)k_\sigma(t)$. We pose the similarity measure as the f-information between the template and the reference image:

$$\mathcal{S}^r(u_T \circ (id + \Phi), u_R) = H_f(u_R) - I_f(u_T \circ (id + \Phi), u_R),$$

and set α and $\mathcal{R}$ as in the previous example. In order to write the derivative of I_f in a compact way, we use the abbreviations $\tilde{\Phi} := id + \Phi$. The derivative of $p_\sigma(u_T \circ \tilde{\Phi})$ with respect to $\tilde{\Phi}$, in direction Ψ, is given by

$$D_{\tilde{\Phi}}(p_\sigma(u_T \circ \tilde{\Phi}), \Psi)(t) = \int_\Omega k'_\sigma(t - u_T(\tilde{\Phi}(x))) \nabla u_T(\tilde{\Phi}(x)) \cdot \Psi(x)\, dx$$

and

$$D_{\tilde{\Phi}}(\hat{p}_\sigma(u_T \circ \tilde{\Phi}, u_R), \Psi)(s,t) = \int_\Omega k'_\sigma(s - u_T(\tilde{\Phi}(x))) \, k_\sigma(t - u_R(x))(\nabla u_T(\tilde{\Phi}(x)) \cdot \Psi(x)) \, dx.$$

We use the following abbreviations:

$$g_1(s,t) := \frac{p_\sigma(u_T \circ \tilde{\Phi})(s) p_\sigma(u_R)(t)}{p_\sigma(u_T \circ \tilde{\Phi}, u_R)(s,t)}, \quad g_2(s,t) := \frac{p_\sigma(u_R)(t)}{(\hat{p}_\sigma(u_T \circ \tilde{\Phi}, u_R)(s,t))^2},$$

and

$$g_3(s,t) := D_{\tilde{\Phi}}(p_\sigma(u_T \circ \tilde{\Phi}), \Psi)(s) \hat{p}_\sigma(u_T \circ \tilde{\Phi}, u_R)(s,t) + p_\sigma(u_T \circ \tilde{\Phi})(s) D_{\tilde{\Phi}}(\hat{p}_\sigma(u_T \circ \tilde{\Phi}, u_R), \Psi)(s,t).$$

With this, we can calculate the derivative of the f-information to be

$$D_{\tilde{\Phi}}(I_f(u_T \circ \tilde{\Phi}, u_R), \Psi) = \int_{\mathbb{R}} \int_{\mathbb{R}} D_{\tilde{\Phi}}(p_\sigma(u_T \circ \tilde{\Phi}), \Psi) \, (s) p_\sigma(u_R)(t) f(g_1(s,t)) + p_\sigma(u_T \circ \tilde{\Phi})(s) p_\sigma(u_R)(t) f'(g_1(s,t)) g_2(s,t) g_3(s,t) \, dt \, ds.$$

For mutual information, this simplifies to

$$D_{\tilde{\Phi}}(MI(u_T \circ \tilde{\Phi}, u_R), \Psi) = \int_{\mathbb{R}} \int_{\mathbb{R}} \left(D_{\tilde{\Phi}} \left(\hat{p}_\sigma \left(u_T \circ \tilde{\Phi}, u_R \right), \Psi \right) \, (s,t) \ln \left(\frac{1}{g_1(s,t)} \right) + \frac{D_{\tilde{\Phi}}(\hat{p}_\sigma(u_T \circ \tilde{\Phi}, u_R), \Psi)(s,t)}{p_\sigma(u_T \circ \tilde{\Phi})(s) p_\sigma(u_R)(t)} \right) ds \, dt.$$

A detailed exposition on elastic registration with mutual information can be found in [9, 11, 16].

In this section, we have presented a general framework on variational-based techniques for nonconstrained multimodal image registration. Below we give a short overview on relevant literature on this topic.

Kim and Fessler [21] describe an intensity based image registration technique that uses a robust correlation coefficient as a similarity measure for images. It is less sensitive to outliers that are present in one image, but not in the other. Kaneko [20] proposed the selective correlation coefficient, as an extension of the correlation coefficient. Van Elsen et al. investigated similarity measures for MR and CT images.

She proposed to calculate the correlation coefficient of geometrical features [37]. Alternatively to the correlation coefficient, one could calculate Spearman's rank correlation coefficient (also known as Spearman's ρ), which is a nonparametric measure of correlation [10], but not very popular in multimodal imaging. Roche et al. [33, 34] tested the correlation ratio to align MR, CT, and PET images. Woods et al. [42] developed an algorithm based on this measure for automated aligning and re-slicing PET images. Independently, several groups realized that the problem of registering two different image modalities can be cast in an information theoretic framework. Collignon et al. [25] and Studholme et al. [35] both suggested using the joint entropy of the combined images as a registration potential. Pluim et al. [28] investigated in more general f-informations. For MR-CT registrations, the learned similarity measure by Lee et al. outperforms all standard measures. Experimental results for learning similarity measures for multimodal images can be found in [22].

5 Recommended Reading

For recent results on divergences and information measures, we refer to *Computational Information Geometry*. Website: http://informationgeometry.wordpress.com/ (last accessed 5 June 2014).

Comparison and evaluation of different similarity measures for CT, MR, and PET brain images can be found in [41].

It is worth mentioning the *Retrospective Image Registration Evaluation Project*. It is designed to compare different multimodal registration techniques. It involves the use of a database of image volumes, commonly known as the "Vanderbilt Database," on which the registrations are to be performed. Moreover, it provides a training data set for multimodal image registration. Link: http://www.insight-journal.org/rire/ (last accessed 5 June 2014). For a collection of databases, we refer to the Validation of Medical Image Registration page http://www.vmip.org (last accessed 5 June 2014).

A number of image registration software tools have been developed in the last decade. The following support multimodal image comparison:

- ITK is an open-source, cross-platform system that provides developers with an extensive suite of software tools for image analysis. It supports the following similarity measures: mean squares metric, normalized cross-correlation metric, mean reciprocal square differences, mutual information (different implementations [26, 39]), the Kullback-Leibler distance, normalized mutual information, correlation coefficient, kappa statistics (for binary images), and gradient difference metric. Website: www.itk.org/ (last accessed 2 June 2014).
- FLIRT is a robust and accurate automated linear (affine) registration tool based around a multi-start, multi-resolution global optimization method. It can be used for inter- and intra-modal registration with 2D or 3D images. Websites: http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FLIRT (last accessed 2 June 2014).

- FAIR, FLIRT are toolboxes for fast and flexible image registration. Both have been developed by the SAFIR-research group in Lübeck. They include the sum of squared differences, mutual information, and normalized gradient fields. Websites: http://www.mic.uni-luebeck.de/de/people/jan-modersitzki/software.html (last accessed 4 June 2014).
- AIR stands for automated image registration. It supports standard deviation of ratio images, least squares, and least squares with global intensity rescaling. Website: http://bishopw.loni.ucla.edu/AIR5/ (last accessed 5 June 2014).
- RView: This software integrates a number of 3D/4D data display and fusion routines together with three-dimensional rigid volume registration using normalized mutual information. It also contains many interactive volume segmentation and painting functions for structural data analysis. Website: http://www.colin-studholme.net/software/software.html (last accessed 5 June 2014).

6 Conclusion

A number of different similarity measures which are used for comparison of multimodal images are listed in this survey. These measures can be grouped into four different categories depending on the way the images are interpreted: If images are represented by pixel intensities (for instance, representing photon counts), one can apply standard l^p, L^p – *norms* for comparison. However, these distance measures are not recommendable for multimodal images. Viewing images as random variables, distance measures from statistics can be utilized. Other common measures are based on image morphology (level lines or gradient fields), which have the advantage that they are less sensitive (or even blind) to gray-level transformations. In applications, images of different modalities neither reveal the same morphology nor the positions of edges are matching. Thus, none of the above mentioned distance measures are useful. For these applications, comparison measures for the gray value distributions of the images were proposed, and in order to obtain estimates of these distributions, kernel density estimations have been implemented on top. Then, one can either compare the distributions of the images directly or alternatively quantify the information content of the joint measures in comparison with the product measures. Such distances are called information measures. Recent approaches are concerned with learning similarity measures with statistical prior information. Variational image registration is discussed, where the appropriate distance measure is important in recovering the right local transformation parameters.

Acknowledgments The work of OS has been supported by the Austrian Science Fund (FWF) within the national research networks Industrial Geometry, project 9203-N12, and Photoacoustic Imaging in Biology and Medicine, project S10505-N20.

The work of CP has been supported by the Austrian Science Fund (FWF) via the Erwin Schrödinger Scholarship J2970.

Cross-References

- Duality and Convex Programming
- Energy Minimization methods
- Non-Linear Image Registration
- Optical Flow
- Regularization Methods for Ill-Posed Problems
- Statistical Methods in Imaging
- Supervised Learning by Support Vector Machines
- Variational Methods in Shape Analysis

References

1. Ali, S.M., Silvey, S.D.: A general class of coefficients of divergence of one distribution from another. J. R. Stat. Soc. B **28**, 131–142 (1966)
2. Bajcsy, R., Kovačič, S.: Multiresolution elastic matching. Comput. Vis. Graph. **46**, 1–21 (1989)
3. Christensen, G.: Deformable shape models for anatomy. PhD thesis, Washington University, Department of Electrical Engineering (1994)
4. Chung, A.C.S., Wells, W.M., Norbash, A., Grimson, W.E.L.: Multi modal image registration by minimizing Kullback-Leibler distance. In: MICCAI'02: Proceedings of the 5th International Conference on Medical Image Computing and Computer-Assisted Intervention-Part II, pp. 525–532. Springer, London (2002)
5. Csiszár, I.: Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten. Magyar Tud Akad Mat Kutató Int Közl **8**, 85–108 (1963)
6. Dragomir, S.S.: Some general divergence measures for probability distributions. Acta Math. Hung. **109**(4), 331–345 (2005)
7. Droske, M., Rumpf, M.: A variational approach to nonrigid morphological image registration. SIAM J. Appl. Math. **64**(2), 668–687 (2003/2004) (electronic)
8. Engl, H.W., Hanke, M., Neubauer, A.: Regularization of Inverse Problems. Mathematics and Its Applications, vol. 375 . Kluwer, Dordrecht (1996)
9. Ens, K., Schumacher, H., Franz, A., Fischer, B.: Improved elastic medical image registration using mutual information. In: Pluim, J.P.W., Reinhardt, J.M. (eds.) Medical Imaging 2007: Image Processing, vol. 6512, p 65122C. SPIE (2007)
10. Fahrmeir, L., Künstler, R., Pigeot, I., Tutz, G.: Statistik, 5th edn. Springer, Berlin (2004)
11. Faugeras, O., Hermosillo, G.: Well-posedness of two nonrigid multi modal image registrational methods. SIAM J. Appl. Math. **64**(5), 1550–1587 (2004). (electronic)
12. Feldman, D., Österreicher, F.: A note on f-divergences. Stud. Sci. Math. Hung. **24**(2–3), 191–200 (1989)
13. Gee, J., Haynor, D., Le Briquer, L., Bajcsyand, R.: Advances in elastic matching theory and its implementation. In: CVRMed-MRCAS'97, Grenoble, pp 63–72, 1997
14. Guetter, C., Xu, C., Sauer, F., Hornegger, J.: Learning based non-rigid multi modal image registration using Kullback-Leibler divergence. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2005, Palm Springs. Lecture Notes in Computer Science, vol. 3750, pp. 255–262. Springer (2005)
15. Haber, E., Modersitzki, J.: Intensity gradient based registration and fusion of multi modal images. In: Methods of Information in Medicine, pp. 726–733. Schattauer Verlag, Stuttgart (2006)

16. Henn, S., Witsch, K.: Multi modal image registration using a variational approach. SIAM J. Sci. Comput. **25**(4), 1429–1447 (2003)
17. Hermosillo, G.: Variational methods for multi modal image matching. PhD thesis, Université de Nice, France (2002)
18. Hömke, L.: A multigrid method for elastic image registration with additional structural constraints. PhD thesis, Heinrich-Heine Universität, Düsseldorf (2006)
19. Ivanov, V.K., Vasin, V.V., Tanana, V.P.: Theory of Linear Ill-Posed Problems and Its Applications. Inverse and Ill-Posed Problems Series, 2nd edn. VSP, Utrecht (2002). (Translated and revised from the 1978 Russian original)
20. Kaneko, S., Satoh, Y., Igarashi, S.: Using selective correlation coefficient for robust image registration. Pattern Recognit. **36**(5), 1165–1173 (2003)
21. Kim, J., Fessler, J.A.: Intensity-based image registration using robust correlation coefficients. IEEE Trans. Med. Imaging **23**(11), 1430–1444 (2004)
22. Lee, D., Hofmann, M., Steinke, F., Altun, Y., Cahill, N.D., Scholkopf, B.: Learning similarity measure for multi modal 3d image registration. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Miami, pp 186–193. IEEE Service Center, Piscataway (2009)
23. Leventon, M.E., Grimson, W.E.L.: Multi modal volume registration using joint intensity distributions. In: MICCAI'98: Proceedings of the First International Conference on Medical Image Computing and Computer-Assisted Intervention, Cambridge, pp. 1057–1066. Springer, London (1998)
24. Likar, B., Pernus, F.: A hierarchical approach to elastic registration based on mutual information. Image Vis. Comput. **19**, 33–44 (2001)
25. Maes, F., Collignon, A., Vandermeulen, D., Marchal, G., Suetens, P.: Multi modality image registration by maximization of mutual information. IEEE Trans. Med. Imaging **16**(2), 187–198 (1997)
26. Mattes, D., Haynor, D.R., Vesselle, H., Lewellen, T.K., Eubank, W.: PET-CT image registration in the chest using free-form deformations. IEEE Trans. Med. Imaging **22**(1), 120–128 (2003)
27. Modersitzki, J.: Numerical Methods for Image Registration. Numerical Mathematics and Scientific Computation. Oxford University Press, Oxford (2004)
28. Pluim, J.P.W., Maintz, J.B.A., Viergever, M.A.: f-information measures in medical image registration. IEEE Trans. Med. Imaging **23**(12), 1508–1516 (2004)
29. Peckar, W., Schnörr, C., Rohr, K., Stiehl, H.S.: Non-rigid image registration using a parameter-free elastic model. In: British Machine Vision Conference, Southampton (1998)
30. Pöschl, C.: Tikhonov regularization with general residual term. PhD thesis, Leopold Franzens Universität, Innsbruck (2008)
31. Rachev, S.T.: Probability Metrics and the Stability of Stochastic Models. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. Wiley, Chichester (1991)
32. Roche, A.: Recalage d'images médicales par inférence statistique. PhD thesis, Université de Nice, Sophia-Antipolis, France (2001)
33. Roche, A., Malandain, G., Pennec, X., Ayache, N.: The correlation ratio as a new similarity measure for multi modal image registration. In: Medical Image Computing and Computer-Assisted Interventation – MICCAI'98, Cambridge. Lecture Notes in Computer Science, vol. 1496, pp. 1115–1124. Springer (1998)
34. Roche, A., Pennec, X., Ayache, N.: The correlation ratio as a new similarity measure for multi modal image registration. In: Medical Image Computing and Computer-Assisted Intervention MICCAI98, Cambridge. Lecture Notes in Computer Science, vol. 1496, pp. 1115–1124. Springer (1998)
35. Studholme, C.: Measures of 3D medical image alignment. PhD thesis, University of London, London (1997)
36. Vajda, I.: Theory of Statistical Inference and Information. Kluwer, Dordrecht (1989)

37. van den Elsen, P., Maintz, J.B.A., Viergever, M.A.: Automatic registration of CT and MR brain images using correlation of geometrical features. IEEE Trans. Med. Imaging. **14**(2), 384–396 (1995)
38. Viola, P.A.: Alignment by maximization of mutual information. PhD thesis, Massachusetts Institute of Technology, Massachusetts (1995)
39. Viola, P.A., Wells, W.M.: Alignment by maximization of mutual information. In: ICCV'95: Proceedings of the Fifth International Conference on Computer Vision, Boston, p. 16. IEEE Computer Society (1995)
40. Wang, X.Y., Feng, D.D.: Automatic elastic medical image registration based on image intensity. Int. J. Image Graph. **5**(2), 351–369 (2005)
41. West, J., Fitzpatrick, J.M., Wang, M.Y., Dawant, B.M., Maurer, C.R., Kessler, R.M., Maciunas, R.J., Barillot, C., Lemoine, D., Collignon, A., Maes, F., Suetens, P., Vandermeulen, D., van den Elsen, P., Napel, S., Sumanaweera, T.S., Harkness, B., Hemler, P.F., Hill, D.L.G., Hawkes, D.J., Studholme, C., Maintz, J.B.A., Viergever, M.A., Malandain, G., Pennec, X., Noz, M.E., Maguire, G.Q., Pollack, M., Pelizzari, C.A., Robb, R.A., Hanson, D., Woods, R.P.: Comparison and evaluation of retrospective intermodality brain image registration techniques. J. Comput. Assist. Tomogr. **21**, 554–566 (1997)
42. Woods, R.P., Cherry, S.R., Mazziotta, J.C.: Rapid automated algorithm for aligning and reslicing PET images. J. Comput. Assist. Tomogr. **16**, 620–633 (1992)

Energy Minimization Methods

Mila Nikolova

Contents

1 Introduction 158
Background 160
The Main Features of the Minimizers as a Function of the Energy 161
Organization of the Chapter 162
2 Preliminaries 162
Notation 162
Reminders and Definitions 163
3 Regularity Results 166
Some General Results 167
Stability of the Minimizers of Energies with Possibly Nonconvex Priors 167
Nonasymptotic Bounds on Minimizers 170
4 Nonconvex Regularization 172
Motivation 172
Assumptions on Potential Functions ϕ 172
How It Works on $\mathbb{R}$ 174
Either Smoothing or Edge Enhancement 175
5 Nonsmooth Regularization 178
Main Theoretical Result 181
Examples and Discussion 183
Applications 185
6 Nonsmooth Data Fidelity 186
General Results 187
Applications 192
7 Nonsmooth Data Fidelity and Regularization 194
The L_1-TV Case 194
ℓ_1 Data Fidelity with Regularization Concave on $\mathbb{R}_+$ 197
8 Conclusion 200
Cross-References 200
References 201

M. Nikolova (✉)
CMLA, ENS Cachan, CNRS, Cachan Cedex, France
e-mail: nikolova@cmla.ens-cachan.fr

© Springer Science+Business Media New York 2015

O. Scherzer (ed.), *Handbook of Mathematical Methods in Imaging*,
DOI 10.1007/978-1-4939-0790-8_5

Abstract

Energy minimization methods are a very popular tool in image and signal processing. This chapter deals with images defined on a discrete finite set. The energies under consideration can be differentiable or not or convex or not. Analytical results on the minimizers of different energies are provided that reveal salient features of the images recovered in this way, as a function of the shape of the energy itself. An intrinsic mutual relationship between energy minimization and modeling via the choice of the energy is thus established. Examples and illustrations corroborate the presented results. Applications that take benefit from these results are presented as well.

1 Introduction

In numerous applications, an unknown image or a signal $u_o \in \mathbb{R}^p$ is represented by data $v \in \mathbb{R}^q$ according to an observation model, called also forward model

$$v = A(u_o) \text{ with noise}, \tag{1}$$

where $A : \mathbb{R}^p \to \mathbb{R}^q$ is a (linear or nonlinear) transform. When u is an $m \times n$ image, its pixels are arranged columnwise into a p-length real vector, where $p = mn$ and the original $u[i, j]$ is identified with $u[(i-1)m + j]$. Some typical applications are, for instance, denoising, deblurring, segmentation, zooming and super-resolution, reconstruction in inverse problems, coding and compression, feature selection, and compressive sensing. In all these cases, recovering a good estimate $\hat{u}$ for u_o needs to combine the observation along with a prior and desiderata on the unknown u_o. A common way to define such an estimate is

$$\text{Find } \hat{u} \text{ such that } \quad \mathcal{F}(\hat{u}, v) = \min_{u \in U} \mathcal{F}(u, v), \tag{2}$$

$$\mathcal{F}(u, v) = \Psi(u, v) + \beta \Phi(u), \tag{3}$$

where $\mathcal{F} : \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}$ is called an energy (or an objective), $U \subset \mathbb{R}^p$ is a set of constraints, Ψ is a data-fidelity term, Φ brings prior information on u_o, and $\beta > 0$ is a parameter which controls the trade-off between Ψ and Φ.

The term Ψ ensures that $\hat{u}$ satisfies (1) quite faithfully according to an appropriate measure. The noise n is random and a natural way to derive Ψ from (1) is to use probabilities; see, e.g., [7, 32, 37, 56]. More precisely, if $\pi(v|u)$ is the likelihood of data v, the usual choice is

$$\Psi(u, v) = -\log \pi(v|u). \tag{4}$$

For instance, if A is a linear operator and $v = Au + n$ where n is additive independent and identically distributed (i.i.d.) zero-mean Gaussian noise, one finds that

$$\Psi(u, v) \propto \|Au - v\|_2^2. \tag{5}$$

This remains quite a common choice partly because it simplifies calculations.

The role of Φ in (3) is to push the solution to exhibit some a priori known or desired features. It is called prior or regularization or penalty term. In many image processing applications, Φ is of the form

$$\Phi(u) = \sum_{i=1}^{r} \phi(\|\mathrm{D}_i u\|), \tag{6}$$

where for any $i \in \{1, \ldots, r\}$, $\mathrm{D}_i : \mathbb{R}^p \to \mathbb{R}^s$, for s an integer $s \geqslant 1$, are linear operators and $\|\cdot\|$ is usually the ℓ_1 or the ℓ_2 norm. For instance, the family $\{\mathrm{D}_i\} \equiv \{\mathrm{D}_i :\in \{1, \ldots, r\}\}$ can represent the discrete approximation of the gradient or the Laplacian operator on u or the finite differences of various orders or the combination of any of these with the synthesis operator of a frame transform or the vectors of the canonical basis of $\mathbb{R}^r$. Note that $s = 1$ if $\{\mathrm{D}_i\}$ are finite differences or a discrete Laplacian; then

$$s = 1 \quad \Rightarrow \quad \phi(\|\mathrm{D}_i u\|) = \phi(|\mathrm{D}_i u|).$$

And if $\{\mathrm{D}_i\}$ are the basis vectors of $\mathbb{R}^r$, one has $\phi(|\mathrm{D}_i u|) = \phi(|u[i]|)$. In (6), $\phi : \mathbb{R}_+ \mapsto \mathbb{R}$ is quite a "general" function, often called a *potential function (PF)*. A very standard assumption is that

H1 *$\phi : \mathbb{R}_+ \to \mathbb{R}$ is proper, lower semicontinuous (l.s.c.) and increasing on $\mathbb{R}_+$, with $\phi(t) > \phi(0)$ for any $t > 0$.*

Some typical examples for ϕ are given in Table 1 and their plots in Fig. 1.

Remark 1. If $\phi'(0^+) > 0$ the function $t \to \phi(|t|)$ is nonsmooth at zero in which case Φ is nonsmooth on $\cup_{i=1}^{r}[w \in \mathbb{R}^p : \mathrm{D}_i w = 0]$. Conversely, $\phi'(0^+) = 0$ leads to a smooth at zero $t \to \phi(|t|)$. With the PF (f13), Φ leads to the counting function, commonly called the ℓ_0-norm.

For the human vision, *an important requirement is that the prior Φ promotes smoothing inside homogeneous regions but preserves sharp edges*. According to a fine analysis conducted in the 1990s and summarized in [7], *ϕ preserves edges if H1 holds as if H2, stated below, holds true* as well:

H2 $\displaystyle\lim_{t \to \infty} \frac{\phi'(t)}{t} = 0.$

This assumption is satisfied by all PFs in Table 1 except for (f1) in case if $\alpha = 2$. Note that there are numerous other heuristics for edge preservation.

Table 1 Commonly used PFs $\phi : \mathbb{R}_+ \to \mathbb{R}$ where $\alpha > 0$ is a parameter. Note that among the *nonconvex PFs*, (f8), (f10), and (f12) are coercive, while the remaining PFs, namely, (f6), (f7), (f9), (f11), and (f13), are bounded. And all nonconvex PFs with $\phi'(0^+) > 0$ are *concave* on $\mathbb{R}_+$. Recall that (f6) is the discrete equivalent of the Mumford-Shah (MS) prior [17, 72]

Convex PFs	
$\phi'(0^+) = 0$	$\phi'(0^+) > 0$
(f1) $\phi(t) = t^\alpha,\ 1 < \alpha \leqslant 2$	(f5) $\phi(t) = t$
(f2) $\phi(t) = \sqrt{\alpha + t^2}$	
(f3) $\phi(t) = \log(\cosh(\alpha t))$	
(f4) $\phi(t) = t/\alpha - \log(1 + t/\alpha)$	
Nonconvex PFs	
$\phi'(0^+) = 0$	$\phi'(0^+) > 0$
(f6) $\phi(t) = \min\{\alpha t^2, 1\}$	(f10) $\phi(t) = t^\alpha,\ 0 < \alpha < 1$
(f7) $\phi(t) = \dfrac{\alpha t^2}{1 + \alpha t^2}$	(f11) $\phi(t) = \dfrac{\alpha t}{1 + \alpha t}$
(f8) $\phi(t) = \log(\alpha t^2 + 1)$	(f12) $\phi(t) = \log(\alpha t + 1)$
(f9) $\phi(t) = 1 - \exp(-\alpha t^2)$	(f13) $\phi(0) = 0,\ \phi(t) = 1$ if $t \neq 0$

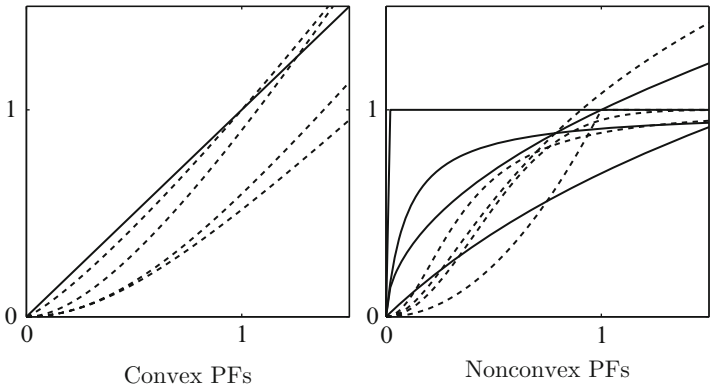


Fig. 1 Plots of the PFs given in Table 1. PFs with $\phi'(0^+) = 0$ (- - -), PFs with $\phi'(0^+) > 0$ (—)

Background

Energy minimization methods, as described here, are *at the crossroad of several well-established methodologies* that are briefly sketched below.

- Bayesian maximum a posteriori (MAP) estimation using Markov random field (MRF) priors. Such an estimation is based on the maximization of the posterior distribution $\pi(u|v) = \pi(v|u)\pi(u)/Z$, where $\pi(u)$ is the prior model for u_o and $Z = \pi(v)$ can be seen as a constant. Equivalently, $\hat{u}$ minimizes with respect to u the energy

$$\mathcal{F}(u, v) = -\ln \pi(v|u) - \ln \pi(u).$$

Identifying these first term above with $\Psi(\cdot, v)$ and the second one with Φ shows the basis of the equivalence. Classical papers on MAP energies using MRF priors are [14–16, 20, 51, 56]. Since the pioneering work of Geman and Geman [56], various *nonconvex* PFs ϕ were explored in order to produce images involving neat edges; see, e.g., [54, 55, 65]. MAP energies involving MRF priors are also considered in many books, such as [32, 53, 64]. For a pedagogical account, see [96].

- Regularization for ill-posed inverse problems was initiated in the book of Tikhonov and Arsenin [93] in 1977. The main idea can be stated in terms of the stabilization of this kind of problems. Useful textbooks in this direction are, e.g., [61, 69, 94] and especially the recent [91]. This methodology and its most recent achievements are nicely discussed from quite a general point of view in Chapter ▸ Regularization Methods for Ill-Posed Problems in this handbook.
- Variational methods are related to PDE restoration methods and are naturally developed for signals and images defined on a continuous subset $\Omega \subset \mathbb{R}^d$, $d = 1, 2, \ldots$; for images $d = 2$. Originally, the data-fidelity term is of the form (5) for $A = \mathrm{Id}$ and $\Phi(u) = \int_\Omega \phi(\|Du\|_2)dx$, where ϕ *is a convex function* as those given in Table 1 (top). Since the beginning of the 1990s, a remarkable effort was done to find heuristics on ϕ that enable *to recover edges and breakpoints in restored images and signals while smoothing the regions between them*; see, e.g., [7, 13, 26, 31, 59, 64, 73, 85, 87]. One of the most successful is the *Total Variation* (TV) regularization corresponding to $\phi(t) = t$, which was proposed by Rudin, Osher, and Fatemi in [87]. Variational methods were rapidly applied along with data-fidelity terms Ψ. The use of differential operators D^k of various orders $k \geqslant 2$ in the prior Φ has been recently investigated; see, e.g., [22, 23]. More details on variational methods for image processing can be found in several textbooks like [3, 7, 91].

  For numerical implementation, the variational functional is discretized and Φ takes the form of (6). Different discretization approaches are considered; see, e.g., [2, 27, 95]

The *equivalence between these approaches* has been considered in several seminal papers; see, e.g., [37, 63]. The state of the art and the relationship among all these methodologies are nicely outlined in the recent book of Scherzer et al. [91]. This book gives a brief historical overview of these methodologies and attaches a great importance to the functional analysis of the presented results.

The Main Features of the Minimizers as a Function of the Energy

Pushing curiosity ahead leads to various additional questions. One observes that frequently data fidelity and priors are modeled separately. It is hence necessary to check if the minimizer $\hat{u}$ of $\mathcal{F}(\cdot, v)$ obeys all information contained in the data model Ψ as well as in the prior Φ. Hence the question: how the prior Φ and the data-fidelity Ψ are *effectively* involved in $\hat{u}$ – a minimizer of $\mathcal{F}(\cdot, v)$. This leads to formulate the following *inverse modeling problem*:

Analyze the mutual relationship between the salient features exhibited by the minimizers $\hat{\boldsymbol{u}}$ of an energy $\mathcal{F}(\cdot, v)$ and the shape of the energy itself. (7)

This problem was posed in a systematic way and studied since [74, 75]. The *point of view* provided by (7) is actually adopted by many authors. Problem (7) is totally general and involves crucial stakes:

- It yields rigorous and strong results on the minimizers $\hat{u}$.
- Such a knowledge enables a real control on the solution – the reconstructed image or signal $\hat{u}$.
- Conversely, it opens *new perspectives for modeling*.
- It enables the conception of specialized energies $\mathcal{F}$ that fulfill the requirements in applications.
- This kind of results can help to derive numerical schemes using knowledge on the solutions.

Problem (7) remains open. The results presented here concern images, signals, and data living on finite grids. In this practical framework, the results in this chapter are quite general since they hold for *energies $\mathcal{F}$ which can be convex or nonconvex or smooth or nonsmooth, and results address local and global minimizers*.

Organization of the Chapter

Some preliminary notions and results that help the reading of the chapter are sketched in Sect. 2. Section 3 is devoted to the regularity of the (local) minimizers of $\mathcal{F}(\cdot, v)$ with a special focus on nonconvex regularization. Section 4 shows how edges are enhanced using nonconvex regularization. In Sect. 5 it is shown that nonsmooth regularization leads typically to minimizers that are sparse in the space spanned by $\{\mathrm{D}_i\}$. Conversely, Sect. 6 exhibits that the minimizers relevant to nonsmooth data fidelity achieve an exact fit for numerous data samples. Section 7 considers results when both Ψ and Φ are nonsmooth. Illustrations and applications are presented.

2 Preliminaries

In this section we set the notations and recall some classical definitions and results on minimization problems.

Notation

We systematically denote by $\hat{u}$ *a (local) minimizer of* $\mathcal{F}(\cdot, v)$. It is explicitly specified when $\hat{u}$ is a global minimizer.

- D_j^n – The differential operator of order n with respect to the jth component of a function.
- $v[i]$ – The ith entry of vector v.
- $\#J$ – The cardinality of the set J.
- $J^c = I \setminus J$ – The complement of $J \subset I$ in I where I is a set.
- $K^\perp$ – The orthogonal complement of a sub-vector space $K \subset \mathbb{R}^n$.
- A^* – The transpose of a matrix (or a vector) where A is real valued.
- $A \succ 0$ ($A \succeq 0$) – The matrix A is positive definite (positive semi-definite)
- $\mathbb{1}_n \in \mathbb{R}^n$ –The n-length vector composed of ones, i.e., $\mathbb{1}_n[i] = 1$, $1 \leqslant i \leqslant n$.
- $\mathbb{L}^n$ – The Lebesgue measure on $\mathbb{R}^n$.
- Id – The identity operator.
- $\|.\|_\rho$ – A vector or a matrix ρ-norm.
- $\mathbb{R}_+ \stackrel{\text{def}}{=} \{t \in \mathbb{R} \,:\, t \geqslant 0\}$ and $\mathbb{R}_+^* \stackrel{\text{def}}{=} \{t \in \mathbb{R} \,:\, t > 0\}$.
- TV – Total Variation.
- $\{e_1, \ldots, e_n\}$ – The canonical basis of $\mathbb{R}^n$, i.e., $e_i[i] = 1$ and $e_i[j] = 0$ if $i \neq j$.

Reminders and Definitions

Definition 1. A function $\mathcal{F} : \mathbb{R}^p \to \mathbb{R}$ is coercive if $\lim_{\|u\| \to \infty} \mathcal{F}(u) = +\infty$.

A special attention being dedicated to nonsmooth functions, we recall some basic facts.

Definition 2. Given $v \in \mathbb{R}^q$, the function $\mathcal{F}(\cdot, v) : \mathbb{R}^p \to \mathbb{R}$ admits at $\hat{u} \in \mathbb{R}^p$ a *one-sided derivative* in a direction $w \in \mathbb{R}^p$, denoted $\delta_1 \mathcal{F}(\hat{u}, v)(w)$, if the following limit exists:

$$\delta_1 \mathcal{F}(\hat{u}, v)(w) = \lim_{t \searrow 0} \frac{\mathcal{F}(\hat{u} + tw, v) - \mathcal{F}(\hat{u}, v)}{t},$$

where the index 1 in δ_1 means that derivatives with respect to the first variable of $\mathcal{F}$ are addressed.

Here $\delta_1 \mathcal{F}(\hat{u}, v)(w)$ is a *right-side* derivative; the *left-side* derivative is $-\delta_1 \mathcal{F}(\hat{u}, v)(-w)$. If $\mathcal{F}(\cdot, v)$ is differentiable at $\hat{u}$, then $\delta_1 \mathcal{F}(\hat{u}, v)(w) = D_1 \mathcal{F}(\hat{u}, v)w$ where D_1 stands for differential with respect to the first variable (see paragraph "Notation"). For $\phi : \mathbb{R}_+ \to \mathbb{R}$, we denote by $\phi'(t^-)$ and $\phi'(t^+)$ its left-side and right-side derivatives, respectively.

The classical necessary condition for a local minimum of a (nonsmooth) function is recalled [60, 86]:

Theorem 1. *If $\mathcal{F}(\cdot, v)$ has a local minimum at $\hat{u} \in \mathbb{R}^p$, then $\delta_1 \mathcal{F}(\hat{u}, v)(w) \geqslant 0$, for every $w \in \mathbb{R}^p$.*

If $\mathcal{F}(\cdot\,, v)$ is Fréchet differentiable at $\hat{u}$, one finds $D_1\mathcal{F}(\hat{u}, v) = 0$.

Rademacher's theorem states that if $\mathcal{F}$ is proper and Lipschitz continuous on $\mathbb{R}^p$, then the set of points in $\mathbb{R}^p$ at which $\mathcal{F}$ is *not* Fréchet differentiable form a set of Lebesgue measure zero [60, 86]. Hence $\mathcal{F}(\cdot\,, v)$ is differentiable at almost every u. However, when $\mathcal{F}(\cdot\,, v)$ is nondifferentiable, its minimizers are typically located at points where $\mathcal{F}(\cdot\,, v)$ is nondifferentiable; see, e.g., Example 1 below.

Example 1. Consider $\mathcal{F}(u, v) = \frac{1}{2}\|u - v\|^2 + \beta|u|$ for $\beta > 0$ and $u, v \in \mathbb{R}$. The minimizer $\hat{u}$ of $\mathcal{F}(\cdot\,, v)$ reads as

$$\hat{u} = \begin{cases} 0 & \text{if } |v| \leqslant \beta \\ v - \operatorname{sign}(v)\beta & \text{if } |v| > \beta \end{cases} \qquad (\hat{u} \text{ is shrunk w.r.t. } v.)$$

Clearly, $\mathcal{F}(\cdot\,, v)$ is not Fréchet differentiable *only at zero*. For any $|v| \leqslant \beta$, the minimizer of $\mathcal{F}(\cdot\,, v)$ is located *precisely at zero.*

The next corollary shows what can happen if the necessary condition in Theorem 1 fails.

Corollary 1. *Let $\mathcal{F}$ be differentiable on $(\mathbb{R}^p \times \mathbb{R}^q) \setminus \Theta_0$ where*

$$\Theta_0 \overset{\text{def}}{=} \{(u, v) \in \mathbb{R}^p \times \mathbb{R}^q : \exists w \in \mathbb{R}^p,\ -\delta_1\mathcal{F}(u, v)(-w) > \delta_1\mathcal{F}(u, v)(w)\}. \tag{8}$$

Given $v \in \mathbb{R}^q$, if $\hat{u}$ is a (local) minimizer of $\mathcal{F}(\cdot\,, v)$ then

$$(\hat{u}, v) \notin \Theta_0.$$

Proof. If $\hat{u}$ is a local minimizer, then by Theorem 1, $\delta_1\mathcal{F}(\hat{u}, v)(-w) \geqslant 0$, hence

$$-\delta_1\mathcal{F}(\hat{u}, v)(-w) \leqslant 0 \leqslant \delta_1\mathcal{F}(\hat{u}, v)(w), \quad \forall w \in \mathbb{R}^p. \tag{9}$$

If $(\hat{u}, v) \in \Theta_0$, the necessary condition (9) cannot hold. □

Example 2. Suppose that Ψ in (3) is a differentiable function for any $v \in \mathbb{R}^q$. For a finite set of positive numbers, say $\theta_1, \ldots, \theta_k$, suppose that the PF ϕ is differentiable on $\mathbb{R}_+ \setminus \cup_{j=1}^k \theta_j$ and that

$$\phi'\left(\theta_j^-\right) > \phi'\left(\theta_j^+\right), \quad 1 \leqslant j \leqslant k. \tag{10}$$

Given a (local) minimizer $\hat{u}$, denote

$$I = \{1, \ldots, r\} \quad \text{and} \quad I_{\hat{u}} = \{i \in I : \|\mathrm{D}_i\hat{u}\|_2 = \theta_j, 1 \leqslant j \leqslant k\}.$$

Define $F(\hat{u}, v) = \Psi(\hat{u}, v) + \beta \sum_{i \in I \setminus I_{\hat{u}}} \phi(\|\mathrm{D}_i \hat{u}\|_2)$, which is differentiable at $\hat{u}$. Clearly, $\mathcal{F}(\hat{u}, v) = F(\hat{u}, v) + \beta \sum_{i \in I_{\hat{u}}} \phi(\|\mathrm{D}_i \hat{u}\|_2)$. Applying the necessary condition (9) for $w = \hat{u}$ yields

$$\beta \sum_{i \in I_{\hat{u}}} \phi' \left(\|\mathrm{D}_i \hat{u}\|_2^- \right) \leqslant -D_1 F(\hat{u}, v)(\hat{u}) \leqslant \beta \sum_{i \in I_{\hat{u}}} \phi' \left(\|\mathrm{D}_i \hat{u}\|_2^+ \right).$$

In particular, one has $\sum_{i \in I_{\hat{u}}} \phi' \left(\|\mathrm{D}_i \hat{u}\|_2^- \right) \leqslant \sum_{i \in I_{\hat{u}}} \phi' \left(\|\mathrm{D}_i \hat{u}\|_2^+ \right)$, which contradicts the assumption on ϕ' in (10). It follows that if $\hat{u}$ is a (local) minimizer of $\mathcal{F}(\cdot, v)$, then $I_{\hat{u}} = \varnothing$ and

$$\|\mathrm{D}_i \hat{u}\|_2 \neq \theta_j, \quad 1 \leqslant j \leqslant k, \quad \forall i \in I.$$

A typical case is the PF (f6) in Table 1, namely, $\phi(t) = \min\{\alpha t^2, 1\}$. Then $k = 1$ and $\theta_1 = \frac{1}{\sqrt{\alpha}}$.

The following existence theorem can be found, e.g., in the textbook [35].

Theorem 2. *For $v \in \mathbb{R}^q$, let $U \subset \mathbb{R}^p$ be a nonempty and closed subset and $\mathcal{F}(\cdot, v) : U \to \mathbb{R}$ a lower semicontinuous (l.s.c.) proper function. If U is unbounded (with possibly $U = \mathbb{R}^p$), suppose that $\mathcal{F}(\cdot, v)$ is coercive. Then there exists $\hat{u} \in U$ such that $\mathcal{F}(\hat{u}, v) = \inf_{u \in U} \mathcal{F}(u, v)$.*

This theorem gives only *sufficient conditions* for the existence of a minimizer. They are *not* necessary, as seen in the example below.

Example 3. Let $\mathcal{F} : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ involve (f6) in Table 1 and read

$$\mathcal{F}(u, v) = (u[1] - v[1])^2 + \beta \phi(|\, u[1] - u[2]\,|) \quad \text{for} \quad \phi(t) = \max\{\alpha t^2, 1\}, \quad 0 < \beta < \infty.$$

For any v, $\mathcal{F}(\cdot, v)$ is not coercive since it is bounded by β in the direction spanned by $\{(0, u[2])\}$. However, its global minimum is strict and is reached for $\hat{u}[1] = \hat{u}[2] = v[1]$ with $\mathcal{F}(\hat{u}, v) = 0$.

To prove the existence of optimal solutions for more general energies, we refer to the textbook [9].

Most of the results summarized in this chapter exhibit the behavior of the minimizer points $\hat{u}$ of $\mathcal{F}(\cdot, v)$ under variations of v. In words, they deal with local minimizer functions.

Definition 3. Let $\mathcal{F} : \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}$ and $O \subseteq \mathbb{R}^q$. We say that $\mathcal{U} : O \to \mathbb{R}^p$ is a *local minimizer function* for the family of functions $\mathcal{F}(\cdot, O) = \{\mathcal{F}(\cdot, v) : v \in O\}$ if for any $v \in O$, the function $\mathcal{F}(\cdot, v)$ reaches a *strict local minimum at* $\mathcal{U}(v)$.

When $\mathcal{F}(\cdot, v)$ is proper, l.s.c., and *convex*, the standard results below can be evoked; see [35, 49].

Theorem 3. *Let $\mathcal{F}(\cdot, v) : \mathbb{R}^p \to \mathbb{R}$ be proper, convex, l.s.c., and coercive for every $v \in \mathbb{R}^q$.*

(i) Then $\mathcal{F}(\cdot, v)$ has a unique (global) minimum which is reached for a closed convex set of minimizers $\left\{\hat{\mathcal{U}}(v)\right\} \stackrel{\text{def}}{=} \left\{\hat{u} \in \mathbb{R}^p : \mathcal{F}(\hat{u}, v) = \inf_{u \in U} \mathcal{F}(u, v)\right\}$.

(ii) If in addition $\mathcal{F}(\cdot, v)$ is strictly convex, then there is a unique minimizer $\hat{u} = \mathcal{U}(v)$ (which is also global). So $\mathcal{F}(\mathbb{R}^p, v)$ has a unique minimizer function $v \mapsto \mathcal{U}(v)$.

The next lemma, which can be found, e.g., in [52], addresses the regularity of the local minimizer functions when $\mathcal{F}$ is smooth. It can be seen as a variant of the implicit functions theorem.

Lemma 1. *Let $\mathcal{F}$ be $\mathcal{C}^m$, $m \geqslant 2$, on a neighborhood of $(\hat{u}, v) \in \mathbb{R}^p \times \mathbb{R}^q$. Suppose that $\mathcal{F}(\cdot, v)$ reaches at $\hat{u}$ a local minimum such that $D_1^2 \mathcal{F}(\hat{u}, v) \succ 0$. Then there are a neighborhood $O \subset \mathbb{R}^q$ containing v and a unique $\mathcal{C}^{m-1}$ local minimizer function $\mathcal{U} : O \to \mathbb{R}^p$, such that $D_1^2 \mathcal{F}(\mathcal{U}(v), v) \succ 0$ for every $v \in O$ and $\mathcal{U}(v) = \hat{u}$.*

This lemma is extended in several directions in this chapter.

Definition 4. Let $\phi : [0, +\infty) \to \mathbb{R}$ and $m \geqslant 0$ an integer. We say that ϕ is $\mathcal{C}^m$ on $\mathbb{R}_+$, or equivalently that $\phi \in \mathcal{C}^m(\mathbb{R}_+)$, if and only if $t \mapsto \phi(|t|)$ is $\mathcal{C}^m$ on $\mathbb{R}$.

By this definition, $\phi'(0) = 0$. In Table 1, left, $\phi \in \mathcal{C}^1(\mathbb{R}_+)$ for (f1) if $\alpha < 2$, $\phi \in \mathcal{C}^2(\mathbb{R}_+)$ for (f4), while for (f2), (f3), and (f7)–(f9) we find $\phi \in \mathcal{C}^\infty(\mathbb{R}_+)$.

3 Regularity Results

Here, we focus on the regularity of the minimizers of $\mathcal{F} : \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}$ of the form

$$\mathcal{F}(u, v) = \|Au - v\|_2^2 + \beta \sum_{i \in I} \phi(\|\mathrm{D}_i u\|_2), \tag{11}$$

$$I \stackrel{\text{def}}{=} \{1, \ldots, r\},$$

where $A \in \mathbb{R}^{q \times p}$ and for any $i \in I$ we have $\mathrm{D}_i \in \mathbb{R}^{s \times p}$ for $s \geqslant 1$. Let us denote by D the following $rs \times p$ matrix:

$$\mathrm{D} \stackrel{\text{def}}{=} \begin{bmatrix} \mathrm{D}_1 \\ \ldots \\ \mathrm{D}_r \end{bmatrix}.$$

When A in (11) is not injective, a standard assumption in order to have regularization is

H3 $\ker(A) \cap \ker(\mathrm{D}) = \{0\}$.

H3 is trivial if $\operatorname{rank} A = p$ or $\operatorname{rank} \mathrm{D} = p$. Often, $\ker(\mathrm{D}) = \operatorname{span}(\mathbb{1}_p)$ and $A\mathbb{1}_p \neq 0$, so H3 holds.

Some General Results

We first verify the conditions on $\mathcal{F}(\cdot, v)$ in (11) that enable Theorems 2 and 3 to be applied. Since H1 holds, $\mathcal{F}(\cdot, v)$ *in* (11) *is l.s.c. and proper.*

1. $\mathcal{F}(\cdot, v)$ in (11) is *coercive* for any $v \in \mathbb{R}^q$ at least in *one* of the following cases:

   - $\operatorname{Rank}(A) = p$ and $\phi : \mathbb{R}_+ \mapsto \mathbb{R}_+$ is nondecreasing.
   - H1 and H3 hold and $\lim_{t \nearrow \infty} \phi(t) = \infty$ (e.g., (f1)–(f5),(f8), (f10), and (f12) in Table 1).

   By Theorem 2, $\mathcal{F}(\cdot, v)$ has minimizers.
2. For any $v \in \mathbb{R}^q$, the energy $\mathcal{F}(\cdot, v)$ in (11) is *convex and coercive* if H1 and H3 hold for a *convex* ϕ. Then the claim in Theorem 3(3) holds true.
3. Further, $\mathcal{F}(\cdot, v)$ in (11) is *strictly convex and coercive* for any $v \in \mathbb{R}^q$ if ϕ satisfies H1 and if *one* of the following assumptions holds:

   - $\operatorname{Rank}(A) = p$ and ϕ is convex.
   - H3 holds and ϕ is strictly convex.

   Then the claim in Theorem 3(3) holds. Further, if $\mathcal{F}$ is $\mathcal{C}^m$ for $m \geqslant 2$, then the minimizer function $\mathcal{U} : \mathbb{R}^q \to \mathbb{R}^p$ (see Definition 3) is $\mathcal{C}^{m-1}$ by Lemma 1.

However, the PFs involved in (11) used for signal and image processing are often nonconvex, bounded, or nondifferentiable. One extension of the standard results is given in the next section.

Stability of the Minimizers of Energies with Possibly Nonconvex Priors

Related questions have been considered in critical point theory, sometimes in semi-definite programming; the well-posedness of some classes of *smooth* optimization problems was addressed in [42]. Other results have been established on the stability of the local minimizers of general *smooth* energies [52]. Typically, these results are quite abstract to be applied directly to energies of the form (11).

Here the assumptions stated below are considered.

H4 *The operator A in (11) satisfies* rank $A = p$, *i.e., A^*A is invertible.*

H5 *The PF ϕ in (11) is $\mathcal{C}^0(\mathbb{R}_+)$ and $\mathcal{C}^m$, $m \geqslant 2$, on $\mathbb{R}_+^*$ with $0 \leqslant \phi'(0^+) < \infty$.*

Under H1, H2, H4, and H5, the prior Φ (and hence $\mathcal{F}(\cdot, v)$) in (11) can be *nonconvex and in addition nonsmooth*. By H1 and H4, $\mathcal{F}(\cdot, v)$ in (11) admits a global minimum $v \in \mathbb{R}^q$ – see Item 1 in section "Some General Results." However, $\mathcal{F}(\cdot, v)$ can present *numerous local minima*.

► Energies $\mathcal{F}$ with nonconvex and possibly nondifferentiable PFs ϕ are frequently used in engineering problems since they were observed to give rise to high-quality solutions $\hat{u}$. It is hence important to have good knowledge on the stability of the obtained solutions.

The results summarized in this section provide the state of the art for energies of the form (11).

Local Minimizers

The stability of local minimizers is an important matter in its own right for several reasons. Often, a nonconvex energy is minimized only locally, in the vicinity of some initial guess. Second, the minimization schemes that guarantee the finding of the global minimum of a nonconvex objective function are exceptional. The practically obtained solutions are usually only local minimizers.

The statements below are a simplified version of the results established in [44].

Theorem 4. *Let $\mathcal{F}(\cdot, v)$ in (11) satisfy H1, H2, H4, and H5. Then there exists a closed subset $\Theta \subset \mathbb{R}^q$ whose Lebesgue measure is $\mathbb{L}^q(\Theta) = 0$ such that for any $v \in \mathbb{R}^q \setminus \Theta$, there exists an open subset $O \subset \mathbb{R}^q$ with $v \in O$ and a local minimizer function (see Definition 3) $\mathcal{U} : O \to \mathbb{R}^p$ which is $\mathcal{C}^{m-1}$ on O and fulfills $\hat{u} = \mathcal{U}(v)$.*

Since Θ is closed in $\mathbb{R}^q$ and $\mathbb{L}^q(\Theta) = 0$, the stated properties are *generic*.

Commentary on the Assumptions

All assumptions H1, H2, and H5 bearing on the PF ϕ are nonrestrictive; they address all PFs in Table 1 except for (f13) which is discontinuous at zero. The assumption H4 cannot be avoided, as seen in Example 4.

Example 4. Consider $\mathcal{F} : \mathbb{R}^2 \times \mathbb{R} \to \mathbb{R}$ given by

$$\mathcal{F}(u, v) = (u[1] - u[2] - v)^2 + |u[1]| + |u[2]|,$$

where $v \equiv v[1]$. The minimum is obtained after a simple computation.

$$v > \frac{1}{2} \qquad \hat{u} = \left(c, c - v + \frac{1}{2}\right) \text{ for any } c \in \left[0, v - \frac{1}{2}\right] \quad \text{(nonstrict minimizer).}$$

$$|v| \leqslant \frac{1}{2} \qquad \hat{u} = 0 \quad \text{(unique minimizer)}$$

$$v < -\frac{1}{2} \qquad \hat{u} = \left(c, c - v - \frac{1}{2}\right) \text{ for any } c \in \left[v + \frac{1}{2}, 0\right] \quad \text{(nonstrict minimizer).}$$

In this case, assumption H4 fails and there is a local minimizer function only for $v \in \left[-\frac{1}{2}, \frac{1}{2}\right]$.

Other Results

The derivations in [44] reveal several other practical results.

1. If $\phi \in \mathcal{C}^2(\mathbb{R}_+)$, see Definition 4, then $\forall v \in \mathbb{R}^q \setminus \Theta$, every local minimizer $\hat{u}$ of $\mathcal{F}(u, v)$ is *strict and* $D_1^2\mathcal{F}(\hat{u}, v) \succ 0$. Consequently, Lemma 1 is extended since the statement holds true $\forall v \in \mathbb{R}^q \setminus \Theta$.

   ► For real data v – a random sample of $\mathbb{R}^q$ – whenever $\mathcal{F}(\cdot, v)$ is differentiable and satisfies the assumptions of Theorem 4, it is a generic property that local minimizers $\hat{u}$ are strict and their Hessians $D_1^2\mathcal{F}(\hat{u}, v)$ are positive definite.

2. Using Corollary 1, the statement of Theorem 4 holds true if $\phi'(0^+) = 0$ and if there is $\tau > 0$ such that $\phi'(\tau^-) > \phi'(\tau^+)$. This is the case of the PF (f6) in Table 1.
3. If $\phi'(0^+) > 0$, define

$$\hat{J} \stackrel{\text{def}}{=} \{i \in I \ : \ \mathrm{D}_i\hat{u} = 0\} \quad \text{and} \quad K_{\hat{J}} \stackrel{\text{def}}{=} \left\{w \in \mathbb{R}^p \ : \ \mathrm{D}_i w = 0, \ \forall i \in \hat{J}\right\}. \tag{12}$$

   Then $\forall v \in \mathbb{R}^q \setminus \Theta$, every local minimizer $\hat{u}$ of $\mathcal{F}(u, v)$ is strict and
   (a) $D_1\mathcal{F}|_{K_{\hat{J}}}(\hat{u}, v) = 0$ and $D_1^2\mathcal{F}|_{K_{\hat{J}}}(\hat{u}, v) \succ 0$ – a sufficient condition for a strict minimum on $K_{\hat{J}}$.
   (b) $\delta_1\mathcal{F}(\hat{u}, v)(w) > 0, \quad \forall w \in K_{\hat{J}}^{\perp} \setminus \{0\}$ – a sufficient condition for a strict minimum on $K_{\hat{J}}^{\perp}$.

   ► Here (a) and (b) provide a sufficient condition for a strict (local) minimum of $\mathcal{F}(\cdot, v)$ at $\hat{u}$ (a direct consequence of [80, Theorem 1]). These conditions are satisfied at the (local) minimizers $\hat{u}$ of $\mathcal{F}(\cdot, v)$ for every $v \in \mathbb{R}^q$, except for a negligible subset of $\mathbb{R}^q$, in which case Lemma 1 can be applied.

   One can interpret these results as follows:

   ► Under the assumptions H1, H2, H4, and H5, given real data $v \in \mathbb{R}^q$, the chance to get a nonstrict (local) minimizer or a (local) minimizer of the energy in (11) that does not result from a $\mathcal{C}^{m-1}$ local minimizer function is null.

Global Minimizers of Energies with for Possibly Nonconvex Priors

The results on the global minimizers of (11) presented next are extracted from [45].

Theorem 5. *Assume that $\mathcal{F}(\cdot, v)$ in (11) satisfy H1, H2, H4, and H5. Then there exists a subset $\hat{\Theta} \subset \mathbb{R}^q$ such that $\mathbb{L}^q(\hat{\Theta}) = 0$ and the interior of $\mathbb{R}^q \setminus \hat{\Theta}$ is dense in $\mathbb{R}^q$, and for any $v \in \mathbb{R}^q \setminus \hat{\Theta}$ the energy $\mathcal{F}(\cdot, v)$ has a unique global minimizer. Furthermore, the global minimizer function $\hat{\mathcal{U}} : \mathbb{R}^q \setminus \hat{\Theta} \to \mathbb{R}^p$ is $\mathcal{C}^{m-1}$ on an open subset of $\mathbb{R}^q \setminus \hat{\Theta}$ which is dense in $\mathbb{R}^q$.*

► Otherwise said, in a real-world problem there is no chance of getting data v such that the energy $\mathcal{F}(\cdot, v)$ (11) has more than one global minimizer.

Nonetheless, $\hat{\Theta}$ plays a crucial role for the recovery of edges; this issue is developed in Sect. 4.

Nonasymptotic Bounds on Minimizers

The aim here is to give *nonasymptotic* analytical bounds on the local and the global minimizers $\hat{u}$ of $\mathcal{F}(\cdot, v)$ in (11) that hold for all PFs ϕ in Table 1. Related questions have mainly been considered in particular cases or asymptotically; see, e.g., [4, 71, 92]. In [51] the mean and the variance of the minimizers $\hat{u}$ for strictly convex and differentiable functions ϕ have been explored.

The bounds provided below are of practical interest for the initialization and the convergence analysis of numerical schemes. The statements given below are extracted from [82].

Bounds on the restored data. One compares the "restored" data $A\hat{u}$ with the given data v.

H6 *Consider the alternative assumptions:*

- $\phi'(0^+) = 0$ *and* $\phi \in \mathcal{C}^1(\mathbb{R}_+ \setminus \Theta_0)$ *where the set* $\Theta_0 = \{t > 0 : \phi'(t^-) > \phi'(t^+)\}$ *is at most finite.*
- $\phi'(0^+) > 0$ *and* ϕ *is* $\mathcal{C}^1$ *on* $\mathbb{R}_+^*$.

The set Θ_0 allows us to address the PF given in (f6). Let us emphasize that under H1 and H6, the PF ϕ can be convex or nonconvex.

Theorem 6. *Consider $\mathcal{F}(\cdot, v)$ of the form (11) where H1, H3, and H6 hold. For every $v \in \mathbb{R}^q$, if $\mathcal{F}(\cdot, v)$ has a (local) minimum at $\hat{u}$, then*

$$\|A\hat{u}\|_2 \leqslant \|v\|_2.$$

Comments on the results. This bound holds for every (local) minimizer of $\mathcal{F}(\cdot, v)$. If A is a uniform tight frame (i.e., $A^*A = \mathrm{Id}$), one has

$$\|\hat{u}\|_2 \leqslant \|v\|_2.$$

The mean of restored data. In many applications, the noise corrupting the data can be supposed to have a mean equal to zero. When $A = \mathrm{Id}$, it is well known that $\mathrm{mean}(\hat{u}) = \mathrm{mean}(v)$; see, e.g., [7]. However, for a general A one has

$$A\mathbb{1}_p \propto \mathbb{1}_q \quad \Rightarrow \quad \mathrm{mean}(\hat{u}) = \mathrm{mean}(v). \tag{13}$$

The requirement $A\mathbb{1}_p \propto \mathbb{1}_q$ is quite restrictive. In the simple case when $\phi(t) = t^2$, $\ker(\mathrm{D}) = \mathbb{1}_{rs}$ and A is square and invertible, it is easy to see that this is also a *sufficient* condition. Finally, if $A \neq \mathrm{Id}$, then generally mean $(\hat{u}) \neq$ mean (v).

The residuals for edge-preserving regularization. A bound on the data-fidelity term at a (local) minimizer $\hat{u}$ of $\mathcal{F}(\cdot, v)$ shall be given. The edge-preserving H2 (see Sect. 1) is replaced by a stronger edge-preserving assumption:

H7 $\quad \|\phi'\|_\infty \stackrel{\mathrm{def}}{=} \max\left\{\sup_{t \geqslant 0} |\phi'(t^+)|,\ \sup_{t>0} |\phi'(t^-)|\right\} < \infty.$

Except for (f1) and (f13), all other PFs in Table 1 satisfy H7. Note that when $\phi'(0^+) > 0$ and H7 hold, one usually has $\|\phi'\|_\infty = \phi'(0^+)$.

Theorem 7. *Let $\mathcal{F}(\cdot, v)$ be of the form (11) where* $\mathrm{rank}\,(A) = q \leqslant p$, *and H1, H3, H6, and H7 hold. For every $v \in \mathbb{R}^q$, if $\mathcal{F}(\cdot, v)$ has a (local) minimum at $\hat{u}$, then*

$$\|A\hat{u} - v\|_\infty \leqslant \frac{\beta}{2}\|\phi'\|_\infty \|(AA^*)^{-1}A\|_\infty \|\mathrm{D}\|_1. \tag{14}$$

Let us emphasize that the bound in (14) is *independent of data v* and that it is satisfied for *any local or global minimizer $\hat{u}$* of $\mathcal{F}(\cdot, v)$. (Recall that for a real matrix C with entries $C[i, j]$, one has $\|C\|_1 = \max_j \sum_i |C[i, j]|$ and $\|C\|_\infty = \max_i \sum_j |C[i, j]|$; see, e.g., [35].)

If D corresponds to a discrete gradient operator for a two-dimensional image, $\|\mathrm{D}\|_1 = 4$. If in addition $A = \mathrm{Id}$, (14) yields

$$\|v - \hat{u}\|_\infty \leqslant 2\beta\|\phi'\|_\infty.$$

The result of this theorem may seem surprising. In a statistical setting, the quadratic data-fidelity term $\|Au - v\|_2^2$ in (11) corresponds to white Gaussian noise on the data, *which is unbounded.* However, if ϕ is edge preserving according to H7, any (local) minimizer $\hat{u}$ of $\mathcal{F}(\cdot, v)$ gives rise to a *noise estimate* $(v - A\hat{u})[i]$, $1 \leqslant i \leqslant q$ *that is tightly bounded* as stated in (14).

► Hence the model for Gaussian noise on the data v is distorted by the solution $\hat{u}$.

► When $\mathcal{F}(\cdot, v)$ is convex and coercive, (14) shows that a good initialization for a minimization algorithm should be a point u_0 such that $Au_0 = v$, e.g., the minimum norm solution of $\|v - \hat{u}\|_2$ given by $u_0 = A^*(AA^*)^{-1}v$.

4 Nonconvex Regularization

Motivation

A permanent requirement is that the energy $\mathcal{F}$ favors the recovery of neat edges. Since the pioneering work of Geman and Geman [56], various *nonconvex* Φ in (3) have been proposed [15,54,55,64,68,72,85]. Indeed, the relevant minimizers exhibit neat edges between homogeneous regions. However, these nonconvex energies are tiresome to control and to minimize (only few algorithms are proved to find the global minimizer in particular cases). In order to avoid these numerical intricacies, since the 1990s, an important effort was done to derive *convex* edge-preserving PFs; see, e.g., [20,31,57,64,87] and [7] for an overview. The most popular convex edge-preserving PF was derived by Rudin, Osher, and Fatemi [87]: it amounts to $\phi = t$, for $\{\mathrm{D}_i\}$ yielding the discrete gradient operator, the ℓ_2-norm in (6) (see Sect. 1), and the relevant Φ is called *the Total Variation (*TV*)* regularization.

In Fig. 2 one sees that the height of the edges is better recovered when ϕ is nonconvex, compared to the convex TV regularization. The same effect can also be observed, e.g., in Figs. 7, 8, and 10.

A considerable progress in nonconvex minimization has been realized. For energies of the form (2)–(3) we refer to [5, 19, 88, 89].

► This section is devoted to explain why edges are nicely recovered using a nonconvex ϕ.

Assumptions on Potential Functions ϕ

Consider $\mathcal{F}(\cdot, v)$ of the form (11) where $\mathrm{D}_i : \mathbb{R}^p \to \mathbb{R}^1$, $i \in I = \{1, \ldots, r\}$, i.e.,

$$\mathcal{F}(u, v) = \|Au - v\|_2^2 + \beta \sum_{i \in I} \phi\left(|\mathrm{D}_i u|\right), \tag{15}$$

and $\phi : \mathbb{R}_+ \to \mathbb{R}_+$ satisfies H1 (see Sect. 1), H6 section "Nonasymptotic Bounds on Minimizers," and H8 given below

H8 *ϕ is $\mathcal{C}^2$ and $\phi'(t) \geqslant 0$ on $\mathbb{R}_+^*$, and $\inf_{t \in \mathbb{R}_+^*} \phi''(t) < 0$ with $\lim_{t \to \infty} \phi''(t) = 0$;*

as well as *one* of the following assumptions:

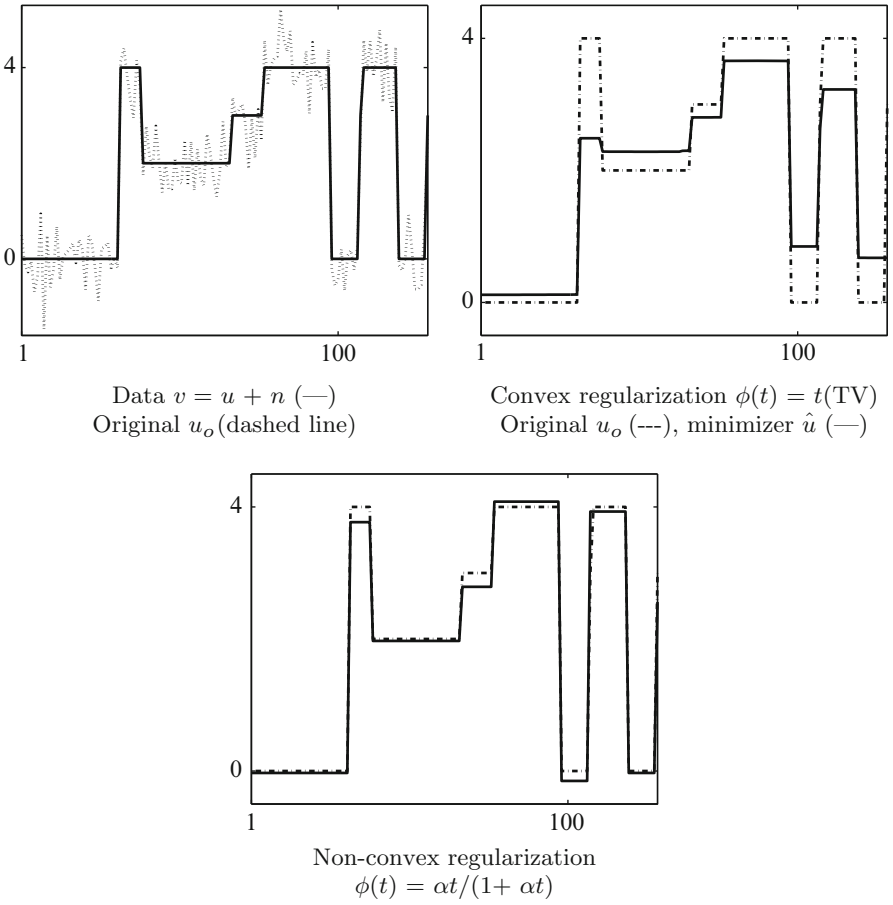


Fig. 2 Minimizers of $\mathcal{F}(u, v) = \|u - v\|_2^2 + \beta \sum_{i=1}^{p-1} \phi(|u[i] - u[i+1]|)$

H9 *$\phi'(0^+) = 0$, and there are two numbers $\tau > 0$ and $\mathcal{T} \in (\tau, \infty)$ such that $\phi''(t) \geqslant 0$ on $[0, \tau]$, $\phi''(t) < 0$ on (τ, ∞), $\phi''(t)$ decreases on $(\tau, \mathcal{T})$ and increases on $(\mathcal{T}, \infty)$.*

H10 *$\phi'(0^+) > 0$, and $\lim_{t \to 0} \phi''(t) < 0$ is well defined and $\phi''(t) < 0$ strictly increases on $(0, \infty)$.*

These assumptions are illustrated in Fig. 3. They hold for all nonconvex PFs in Table 1, except for (f6) and (f13) which are presented separately. Further, these assumptions are easy to relax.

The results presented below come from [81].

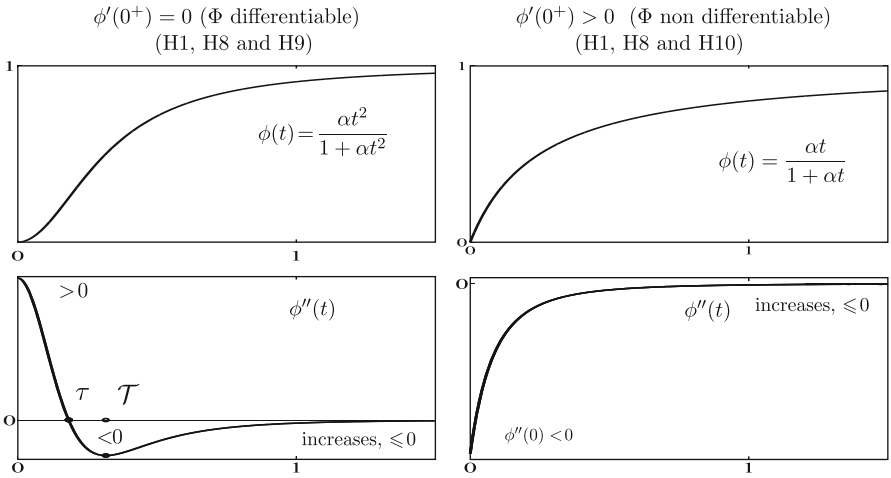


Fig. 3 Illustration of the assumptions in two typical cases – (f7) and (f11) – in Table 1

How It Works on $\mathbb{R}$

► This example illustrates the main facts that explain why edges are enhanced when ϕ is nonconvex, satisfying H1, and H8 along with either H9 or H10.

Let $\mathcal{F} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be given by

$$\mathcal{F}(u, v) = \frac{1}{2}(u - v)^2 + \beta\phi(u) \text{ for } \begin{cases} \beta > \frac{1}{|\phi''(\mathcal{T})|} & \text{if } \phi'(0^+) = 0 \text{ (H1, H8 and H9)} \\ \beta > \frac{1}{\left|\lim_{t \searrow 0} \phi''(t)\right|} & \text{if } \phi'(0^+) > 0 \text{ (H1, H8 and H10)} \end{cases}$$

The (local) minimality conditions for $\hat{u}$ of $\mathcal{F}(\cdot, v)$ read as

- If $\phi'(0^+) = 0$ or $\left[\phi'(0^+) > 0 \text{ and } \hat{u} \neq 0\right] : \hat{u} + \beta\phi'(\hat{u}) = v$ and $1 + \beta\phi''(\hat{u}) \geqslant 0$.
- If $\phi'(0^+) > 0$ and $\hat{u} = 0 : |v| \leqslant \beta\phi'(0^+)$.

To simplify, we assume that $v \geqslant 0$. Define

$$\theta_0 = \inf C_\beta \text{ and } \theta_1 = \sup C_\beta,$$
$$\text{for } C_\beta = \left\{u \in \mathbb{R}_+^* : D_1^2\mathcal{F}(u, v) < 0\right\} = \left\{u \in \mathbb{R}_+^* : \phi''(u) < -1/\beta\right\}.$$

One has $\theta_0 = 0$ if $\phi'(0^+) > 0$ and $0 < \theta_0 < \mathcal{T} < \theta_1$ if $\phi'(0^+) = 0$. A few calculations yield

1. For every $v \in \mathbb{R}_+$ *no minimizer lives in* (θ_0, θ_1) (cf. Fig. 4).
2. One computes $0 < \xi_0 < \xi_1$ such that (cf. Fig. 4)
   a. If $0 \leqslant v \leqslant \xi_1$, $\mathcal{F}(\cdot, v)$ has a (local) minimizer $\hat{u}_0 \in [0, \theta_0]$, hence $\hat{u}_0$ is subject to a strong smoothing.
   b. If $v \geqslant \xi_0$, $\mathcal{F}(\cdot, v)$ has a (local) minimizer $\hat{u}_1 \geqslant \theta_1$, hence $\hat{u}_1$ is subject to a weak smoothing.
   c. If $v \in [\xi_0, \xi_1]$ then $\mathcal{F}(\cdot, v)$ has two local minimizers, $\hat{u}_0$ and $\hat{u}_1$.
3. There is $\xi \in (\xi_0, \xi_1)$ such that $\mathcal{F}(\cdot, \xi)$ has two global minimizers, $\mathcal{F}(\hat{u}_0, \xi) = \mathcal{F}(\hat{u}_1, \xi)$, as seen in Fig. 5;
   a. If $0 < v < \xi$, the unique global minimizer is $\hat{u} = \hat{u}_0$.
   b. If $v > \xi$, the unique global minimizer is $\hat{u} = \hat{u}_1$.
4. The global minimizer function $v \mapsto \mathcal{U}(v)$ is discontinuous at ξ and $\mathcal{C}^1$-smooth on $\mathbb{R}_+ \setminus \{\xi\}$.

Item 1 is the key for the recovery of either homogeneous regions or high edges. The minimizer $\hat{u}_0$ (see Items 2a and 3a) corresponds to the restoration of homogeneous regions, while $\hat{u}_1$ (see Items 2b and 3b) corresponds to edges. Item 3 corresponds to a decision for the presence of an edge at the global minimizer. Since $\{\xi\}$ is closed and $\mathbb{L}^1\{\xi\} = 0$, Item 4 confirms the results of section "Global Minimizers of Energies with for Possibly Nonconvex Priors."

Either Smoothing or Edge Enhancement

(A) Case $\phi'(0^+) = 0$. Below the case depicted in Figs. 4, left, and 5, left, is extended to $\mathbb{R}^p$.

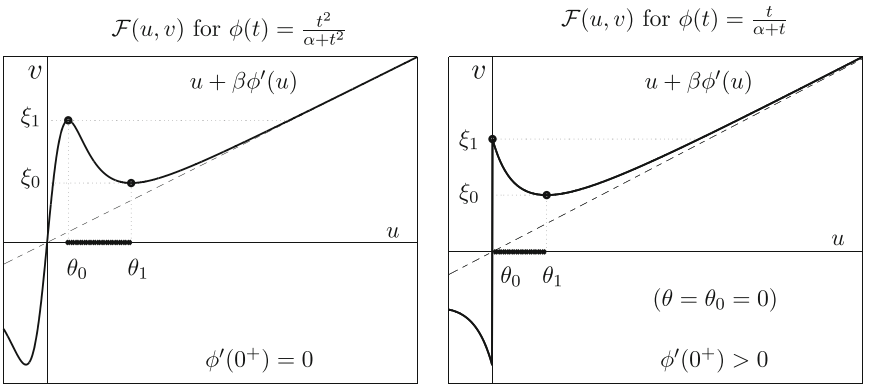

$\mathcal{F}(u, v)$ for $\phi(t) = \frac{t^2}{\alpha + t^2}$
$\mathcal{F}(u, v)$ for $\phi(t) = \frac{t}{\alpha + t}$
v
$u + \beta\phi'(u)$
ξ_1
ξ_0
u
θ_0
θ_1
$\phi'(0^+) = 0$
$(\theta = \theta_0 = 0)$
$\phi'(0^+) > 0$
Minimizer if : $\hat{u} + \beta\phi'(\hat{u}) = v$ and $1 + \beta\phi''(\hat{u}) > 0$
Minimizer at 0 : $|v| \leqslant \beta\phi'(0^+)$ Else : $\hat{u} + \beta\phi'(\hat{u}) = v$ and $1 + \beta\phi''(\hat{u}) > 0$

Fig. 4 The curve of $u \mapsto \big(D_1\mathcal{F}(u, v) - v\big)$ on $\mathbb{R} \setminus \{0\}$. All assumptions mentioned before hold

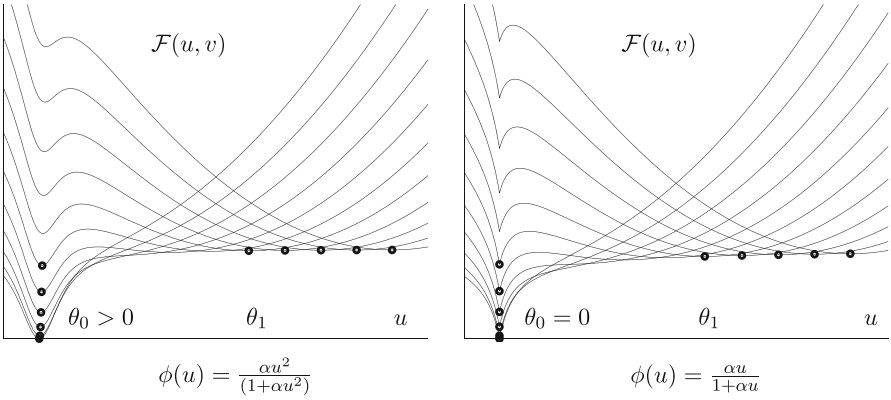


Fig. 5 Each curve represents $\mathcal{F}(u,v) = \frac{1}{2}(u-v)^2 + \beta\phi(u)$ for an increasing sequence $v \in [0, \xi_1)$. The global minimizer of each $\mathcal{F}(\cdot\,, v)$ is marked with "•." No (local) minimizer lives in (θ_0, θ_1)

Theorem 8. *Let $\mathcal{F}(\cdot\,, v)$ be of the form (15) where H1, H3, H8, and H9 hold, and $\{\mathrm{D}_i : i \in I\}$ are linearly independent. Set $\mu \overset{\text{def}}{=} \max_{1\leqslant i\leqslant r} \|\mathrm{D}^*(\mathrm{D}\mathrm{D}^*)^{-1}e_i\|_2$. For $\beta > \frac{2\mu^2\,\|A^*A\|_2}{|\phi''(\mathcal{T})|}$, there are $\theta_0 \in (\tau, \mathcal{T})$ and $\theta_1 \in (\mathcal{T}, \infty)$ such that $\forall v \in \mathbb{R}^q$, if $\hat{u}$ is a (local) minimizer of $\mathcal{F}(\cdot\,, v)$, then*

$$\textit{either} \quad |\mathrm{D}_i\hat{u}| \leqslant \theta_0, \quad \textit{or} \quad |\mathrm{D}_i\hat{u}| \geqslant \theta_1, \quad \forall i \in I. \tag{16}$$

In many imaging problems, $\{\mathrm{D}_i\}$ are not linearly independents. If $\{\mathrm{D}_i\}$ are linearly dependent, the result (16) holds true for all (local) minimizers $\hat{u}$ that are locally homogeneous on regions that are connected with respect to $\{\mathrm{D}_i\}$. Otherwise, one recovers both high edges and smooth transitions, as seen in Fig. 8a. When ϕ is convex, all edges are smoothed, as one can observe in Fig. 7a.

The PF $\phi(t) = \min\{\alpha t^2, 1\}$ (f6) in Table 1 does not satisfy assumptions H8 and H9. From Corollary 1 and Example 2 (section "Reminders and Definitions"), any (local) minimizer $\hat{u}$ of $\mathcal{F}(\cdot\,, v)$ obeys

$$|\mathrm{D}_i\hat{u}| \neq \frac{1}{\sqrt{\alpha}}, \quad \forall i \in I.$$

Proposition 1 below addresses only the *global minimizers* of $\mathcal{F}(\cdot\,, v)$.

Proposition 1. *Let $\mathcal{F}(\cdot\,, v)$ be given by (15) where $\phi(t) = \min\{\alpha t^2, 1\}$, $\{\mathrm{D} : i \in I\}$ are linearly independent and $\mathrm{rank}\,(A) \geqslant p - r \geqslant 1$. If $\mathcal{F}(\cdot\,, v)$ has a global minimizer at $\hat{u}$, then*

$$\textit{either} \quad |\mathrm{D}_i\hat{u}| \leqslant \frac{1}{\sqrt{\alpha}}\,\Gamma_i, \quad \textit{or} \quad |\mathrm{D}_i\hat{u}| \geqslant \frac{1}{\sqrt{\alpha}\,\Gamma_i},$$

$$\textit{for} \quad \Gamma_i = \sqrt{\frac{\|Be_i\|_2^2}{\|Be_i\|_2^2 + \alpha\beta}} < 1, \quad \forall i \in I, \tag{17}$$

where B is a matrix depending only on A and D.

If $\mathrm{D} = \mathrm{Id}$, then $B = A$. If u one-dimensional signal and $\mathrm{D}_i u = u[i] - u[i+1]$, $1 \leqslant i \leqslant p-1$, one has $B = (\mathrm{Id} - \frac{1}{p}\mathbb{1}\mathbb{1}^T)AH$ where $H \in \mathbb{R}^{p\times p}$ is upper triangular composed of ones.

In Proposition 1, set $\theta_0 = \dfrac{\gamma}{\sqrt{\alpha}}$ and $\theta_1 = \dfrac{1}{\sqrt{\alpha}\gamma}$ for $\gamma \overset{\text{def}}{=} \max\limits_{i\in I} \Gamma_i < 1$.

Let us define the following subsets:

$$\hat{J}_0 \overset{\text{def}}{=} \{i \in I \ : \ |\mathrm{D}_i \hat{u}| \leqslant \theta_0\} \quad \text{and} \quad \hat{J}_1 \overset{\text{def}}{=} I \setminus \hat{J}_0 = \{i \in I \ : \ |\mathrm{D}_i \hat{u}| \geqslant \theta_1\}. \tag{18}$$

One can interpret the results of Theorem 8 and Proposition 1 as follows:

- The pixels in $\hat{J}_0$ form homogeneous regions with respect to $\{\mathrm{D}_i\}$, whereas the pixels in $\hat{J}_1$ are break points.

  In particular, if $\{\mathrm{D}_i\}$ correspond to first-order differences, $\hat{J}_0$ addresses smoothly varying regions where $|\mathrm{D}_i \hat{u}| \leqslant \theta_0$, while $\hat{J}_1$ corresponds to edges higher than $\theta_1 - \theta_0$.

(B) Case $\boldsymbol{\phi'(0^+) > 0}$. Here the results are stronger without assumptions on $\{\mathrm{D}_i\}$. This case corresponds to Figs. 4, right, and 5, right.

Theorem 9. *Consider $\mathcal{F}(\cdot, v)$ of the form (15) where H3 holds and ϕ satisfies H1, H8, and H10. Let $\beta > \frac{2\mu^2 \|A^*A\|_2}{|\lim_{t\searrow 0} \phi''(t)|}$, where $\mu > 0$ is a constant depending only on $\{\mathrm{D}_i\}$. Then $\exists \theta_1 > 0$ such that $\forall v \in \mathbb{R}^q$, every (local) minimizer $\hat{u}$ of $\mathcal{F}(\cdot, v)$ satisfies*

$$\textit{either} \quad |\mathrm{D}_i \hat{u}| = 0, \quad \textit{or} \quad |\mathrm{D}_i \hat{u}| \geqslant \theta_1, \quad \forall i \in I. \tag{19}$$

The results of Theorem 9 were extended to energies involving box constraints in [33].

The "0-1" PF (f13) in Table 1 does not satisfy H8 and H10 since it is discontinuous at 0.

Proposition 2. *Let $\mathcal{F}(\cdot, v)$ in (15) be defined for $\phi(0) = 0$, $\phi(t) = 1$ if $t > 0$, i.e., (f13), $\{\mathrm{D}_i\}$ be linearly independent and* rank $A \geqslant p - r \geqslant 1$. *If $\mathcal{F}(\cdot, v)$ has a* global minimum *at $\hat{u}$, then*

$$\textit{either} \quad |\mathrm{D}_i \hat{u}| = 0 \quad \textit{or} \quad |\mathrm{D}_i \hat{u}| \geqslant \frac{\sqrt{\beta}}{\|Be_i\|_2}, \quad \forall i \in I, \tag{20}$$

where the matrix B depends only on D *and on A.*

In (20), B is the same as in Proposition 1. For $\theta_1 \stackrel{\text{def}}{=} \min_{i\in I} \frac{\sqrt{\beta}}{\|Be_i\|}$, it is clear that (20) holds.

Let

$$\hat{J}_0 \stackrel{\text{def}}{=} \{i \ : \ |\mathrm{D}_i \hat{u}| = 0\} \text{ and } \hat{J}_1 \stackrel{\text{def}}{=} I \setminus \hat{J}_0 = \{i \ : \ |\mathrm{D}_i \hat{u}| \geqslant \theta_1\} \, .$$

Using this notation, the results of Theorem 9 and Proposition 2 show that:

► The indexes in $\hat{J}_0$ address regions in $\hat{u}$ that can be called *strongly homogeneous* (since $|\mathrm{D}_i \hat{u}| = 0$), while $\hat{J}_1$ addresses breakpoints where $|\mathrm{D}_i \hat{u}| \geqslant \theta_1$.

If $\{\mathrm{D}_i\}$ are first-order differences, $\hat{u}$ is *neatly segmented*: $\hat{J}_0$ corresponds to *constant* regions, while $\hat{J}_1$ describes all edges and they are higher than θ_1.

Direct segmentation of an image from data transformed via a general (nondiagonal) operator A remains a difficult task using standard methods. The result in (19), Theorem 9, tells us that such a segmentation is naturally involved in the minimizers $\hat{u}$ of $\mathcal{F}(\cdot, v)$, for *any operator* A. This effect can be observed, e.g., on Figs. 8b, d and 11d.

(C) Illustration: Deblurring of an image from noisy data. The original image u_o in Fig. 6a presents smoothly varying regions, constant regions, and sharp edges. Data in Fig. 6b correspond to $v = a * u_o + n$, where a is a blur with entries $a_{i,j} = \exp\left(-(i^2 + j^2)/12.5\right)$ for $-4 \leqslant i, j \leqslant 4$, and n is white Gaussian noise yielding 20 dB of SNR. The amplitudes of the original image are in the range of $[0, 1.32]$ and those of the data in $[-5, 50]$. In all restored images, $\{\mathrm{D}_i\}$ correspond to the first-order differences of each pixel with its 8 nearest neighbors. In all figures, the obtained minimizers are displayed on the top. Just below, the sections corresponding to rows 54 and 90 of the restored images are compared with the same rows of the original image.

The restorations in Fig. 7 are obtained using *convex PFs* ϕ while those in Fig. 8 using *nonconvex PFs* ϕ. Edges are sharp and high in Fig. 8 where ϕ is nonconvex, which corroborates the results in paragraphs (A) and (B). In Fig. 8b, d ϕ *is nonconvex and* $\phi'(0^+) > 0$ *in addition*. As stated in Theorem 9, in spite of the fact that A is nondiagonal (and ill-conditioned), the restored images are fully segmented and the edges between constant pieces are high.

5 Nonsmooth Regularization

► Observe that the minimizers corresponding to $\phi'(0^+) > 0$ (nonsmooth regularization) in Figs. 2b, c, 7b, 8b, d, 10a–c, and 11d are constant on numerous regions. This section is aimed to explain and to generalize this observation.

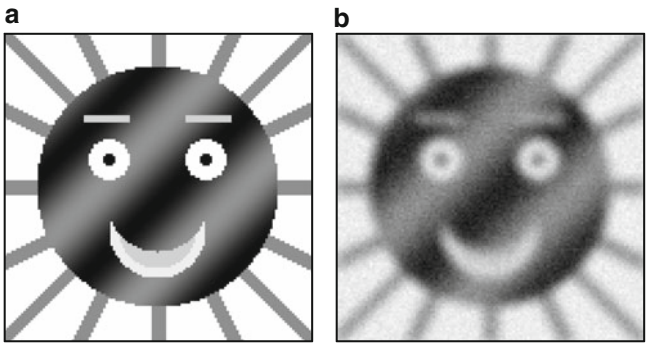


Fig. 6 Data $v = a \star u_o + n$, where a is a blur and n is white Gaussian noise, 20 dB of SNR. (**a**) Original image. (**b**) Data $v =$ blur + noise

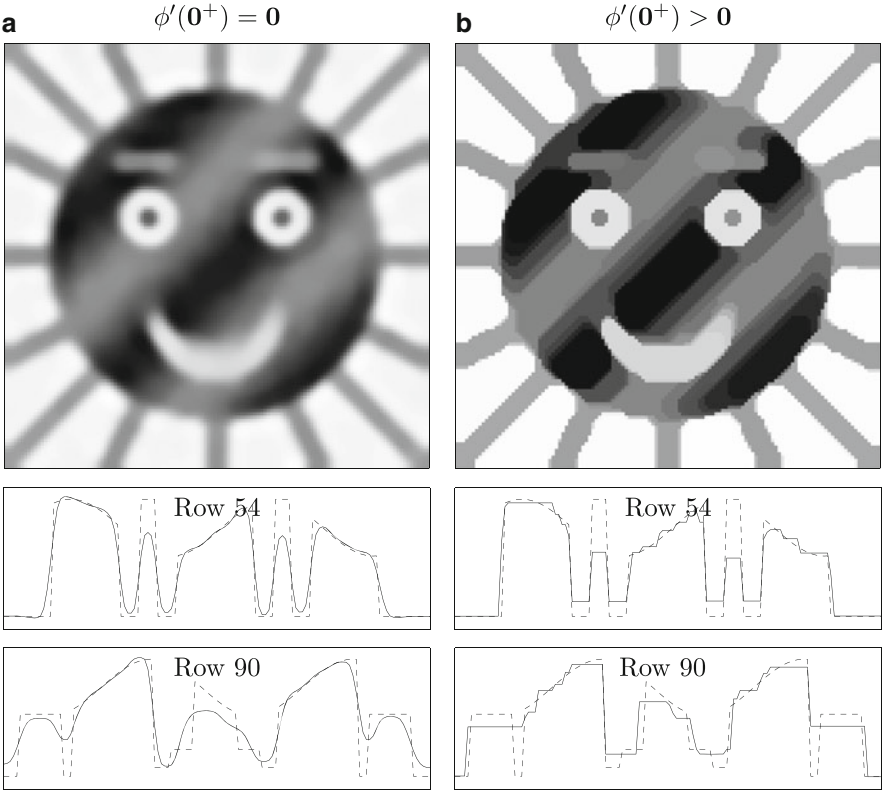


Fig. 7 Restoration using convex PFs. (**a**) $\phi(t) = t^\alpha$ for $\alpha = 1.4, \beta = 40$. (**b**) $\phi(t) = t$ for $\beta = 100$

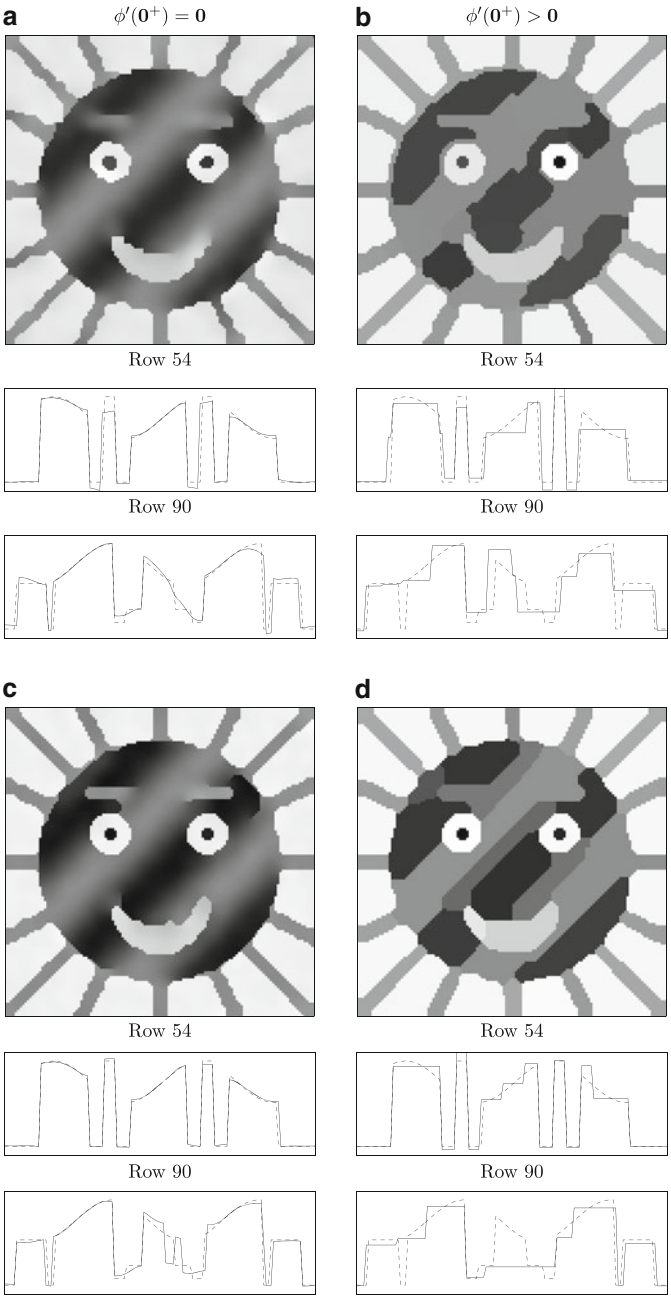


Fig. 8 Restoration using nonconvex PFs. (**a**) $\phi(t) = \dfrac{\alpha t^2}{1+\alpha t^2}$ for $\alpha = 25$, $\beta = 35$. (**b**) $\phi(t) = \dfrac{\alpha t}{1+\alpha t}$ for $\alpha = 20$, $\beta = 100$. (**c**) $\phi(t) = \min\{\alpha t^2, 1\}$ for $\alpha = 60$, $\beta = 10$. (**d**) $\phi(0) = 0$, $\phi(t) = 1$, $t > 0$ for $\beta = 25$

Consider

$$\mathcal{F}(u,v) = \Psi(u,v) + \beta\Phi(u) \tag{21}$$

$$\Phi(u) = \sum_{i=1}^{r} \phi\left(\left\|\mathrm{D}_i u\right\|_2\right), \tag{22}$$

where $\Psi : \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}$ is *any explicit or implicit* $\mathcal{C}^m$-smooth function for $m \geqslant 2$ and $\mathrm{D}_i : \mathbb{R}^p \mapsto \mathbb{R}^s$, $\forall i \in I = \{1, \ldots, r\}$, are general linear operators for any integer $s \geqslant 1$. It is assumed that ϕ satisfies H1 along with

H11 *ϕ is $\mathcal{C}^2$-smooth on $\mathbb{R}_+^*$ and $\phi'(0^+) > 0$.*

Note that Ψ *and ϕ can be convex or nonconvex*. Let us define the set-valued function $\mathcal{J}$ on $\mathbb{R}^p$ by

$$\mathcal{J}(u) = \left\{i \in I \,:\, \|\mathrm{D}_i u\|_2 = 0\right\}. \tag{23}$$

Given $u \in \mathbb{R}^p$, $\mathcal{J}(u)$ indicates all regions where $\mathrm{D}_i u = 0$. Such regions are called *strongly homogeneous* with respect to $\{\mathrm{D}_i\}$. (The adverb "strongly" is used to emphasize the difference with just "homogeneous regions" where $\|\mathrm{D}_i u\|_2 \approx 0$.) In particular, if $\{\mathrm{D}_i\}$ correspond to first-order differences between neighboring samples of u or to discrete gradients, $\mathcal{J}(u)$ indicates all constant regions in u.

Main Theoretical Result

The results presented below are extracted from [80].

Theorem 10. *Given $v \in \mathbb{R}^q$, assume that $\mathcal{F}(\cdot, v)$ in (21)–(22) is such that Ψ is $\mathcal{C}^m$, $m \geqslant 2$ on $\mathbb{R}^p \times \mathbb{R}^q$, and that ϕ satisfies H1 and H11. Let $\hat{u} \in \mathbb{R}^p$ be a (local) minimizer of $\mathcal{F}(\cdot, v)$. For $\hat{J} \stackrel{\text{def}}{=} \mathcal{J}(\hat{u})$, let $K_{\hat{J}}$ be the vector subspace*

$$K_{\hat{J}} = \left\{u \in \mathbb{R}^p : \mathrm{D}_i u = 0, \forall i \in \hat{J}\right\}. \tag{24}$$

Suppose also that

(a) $\delta_1\mathcal{F}(\hat{u}, v)(w) > 0$, *for every* $w \in K_{\hat{J}}^{\perp} \setminus \{0\}$.

(b) *There is an open subset $O'_{\hat{J}} \subset \mathbb{R}^q$ such that $\mathcal{F}|_{K_{\hat{J}}}\left(., O'_{\hat{J}}\right)$ has a local minimizer function $\mathcal{U}_{\hat{J}} : O'_{\hat{J}} \to K_{\hat{J}}$ which is $\mathcal{C}^{m-1}$ continuous and $\hat{u} = \mathcal{U}_{\hat{J}}(v)$.*

Then there is an open neighborhood $O_{\hat{J}} \subset O'_{\hat{J}}$ of v such that $\mathcal{F}(\cdot, O_{\hat{J}})$ admits a $\mathcal{C}^{m-1}$ local minimizer function $\mathcal{U} : O_{\hat{J}} \to \mathbb{R}^p$ which satisfies $\mathcal{U}(v) = \hat{u}$, $\mathcal{U}|_{K_{\hat{J}}} = \mathcal{U}_{\hat{J}}$ and

$$v \in O_{\hat{J}} \;\Rightarrow\; \mathrm{D}_i \mathcal{U}(v) = 0, \textit{ for all } i \in \hat{J}. \tag{25}$$

Note that $\hat{J}$ and $K_{\hat{J}}$ are the same as those introduced in (12) section "Local Minimizers."

Commentary on the assumptions. Since $\mathcal{F}(\cdot, v)$ has a local minimum at $\hat{u}$, by Theorem 1 one has $\delta_1 \mathcal{F}(\hat{u}, v)(w) \geqslant 0$, for all $w \in K_{\hat{J}}^{\perp} \setminus \{0\}$, and if for some w the inequality becomes inequality, then the inequality is strict for $-w$. So (a) is not a strong requirement. Condition (b) amounts to Lemma 1 (section "Reminders and Definitions") applied to $\mathcal{F}|_{K_{\hat{J}}}$ which is $\mathcal{C}^m$ on a neighborhood of $(\hat{u}, v)$ belonging to $K_{\hat{J}} \times \mathbb{R}^q$.

If $\mathcal{F}(\cdot, v)$ (possibly nonconvex) is of the form (11) and assumption H4 (section "Stability of the Minimizers of Energies with Possibly Nonconvex Priors") holds, Theorem 4 and the other results given next show that (a) and (b) are satisfied for any $v \in \mathbb{R}^q \setminus \Theta$ where Θ is closed and $\mathbb{L}^q(\Theta) = 0$.

Significance of the results. Using the definition of $\mathcal{J}$ in (23), the conclusion of the theorem can be reformulated as

$$v \in O_{\hat{J}} \;\Rightarrow\; \mathcal{J}\left(\mathcal{U}(v)\right) \supseteq \hat{J} \;\Leftrightarrow\; \mathcal{U}(v) \in K_{\hat{J}}. \tag{26}$$

Minimizers involving large subsets $\hat{J}$ are observed in Figs. 2b, c, 7b, 8b, d, 10a–c, and 11d. It was seen in Examples 1 and 4, as well as in section "How It Works on $\mathbb{R}$" (case $\phi'(0^+) > 0$), that $\hat{J}$ is nonempty for data v living in an open $O_{\hat{J}}$. Note also that there is an open subset $\tilde{O}_{\hat{J}} \subset O_{\hat{J}}$ such that $\mathcal{J}\left(\mathcal{U}(v)\right) = \hat{J}$ for all $v \in \tilde{O}_{\hat{J}}$. These sets $\tilde{O}_{\hat{J}}$ are described in Example 5.

Observe that (26) is a severe restriction since $K_{\hat{J}}$ is a closed and negligible subset of $\mathbb{R}^p$, whereas data v vary on open subsets $O_{\hat{J}}$ of $\mathbb{R}^q$.

Focus on a (local) minimizer function $\mathcal{U} : O \to \mathbb{R}^p$ for $\mathcal{F}(\cdot, O)$ and put $\hat{J} = \mathcal{J}(\mathcal{U}(v))$ for some $v \in O$. By Theorem 10, the sets $O_{\hat{J}}$ and $\tilde{O}_{\hat{J}}$ are of positive measure in $\mathbb{R}^q$. When data v range over O, the set-valued function $(\mathcal{J} \circ \mathcal{U})$ generally takes several distinct values, say $\{J_j\}$. Thus, with a (local) minimizer function $\mathcal{U}$, defined on an open set O, there is associated a family of subsets $\{\tilde{O}_{J_j}\}$ which form a covering of O. When $v \in \tilde{O}_{J_j}$, we find a minimizer $\hat{u} = \mathcal{U}(v)$ satisfying $\mathcal{J}(\hat{u}) = J_j$.

► Energies with nonsmooth regularization terms, as those considered here, exhibit local minimizers which *generically* satisfy constraints of the form $\mathcal{J}(\hat{u}) \neq \varnothing$.

In particular, if $\{\mathrm{D}_i\}$ are discrete gradients or first-order difference operators, minimizers $\hat{u}$ are typically constant on many regions. For example, if $\phi(t) = t$, we have $\Phi(u) = \mathrm{TV}(u)$, and this explains the stair-casing effect observed in TV methods on discrete images and signals [30, 39].

Examples and Discussion

The subsection begins with an illustration of Theorem 10 and its meaning.

Restoration of a noisy signal. Figure 9 shows a piecewise constant signal u_o corrupted with two different noises.

Figure 10 depicts the restoration from these two noisy data samples by minimizing an energy of the form $\mathcal{F}(u, v) = \|u - v\|^2 + \beta \sum_{i=1}^{p-1} \phi(|u[i] - u[i+1]|)$. The minimizers shown in Fig. 10a–c correspond to functions ϕ such that $\phi'(0^+) > 0$ and they are constant on large segments. The reader is invited to compare the subsets where these minimizers are constant. The function ϕ in Fig. 10d satisfies $\phi'(0^+) = 0$ and the resultant minimizers are nowhere constant.

Example 5 below gives a rich geometric interpretation of Theorem 25.

Example 5 (1D TV Regularization). Let $\mathcal{F} : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ be given by

$$\mathcal{F}(u, v) = \|Au - v\|_2^2 + \beta \sum_{i=1}^{p-1} \big|u[i] - u[i+1]\big|, \quad \beta > 0, \tag{27}$$

where $A \in \mathbb{R}^{p \times p}$ is invertible. Clearly, there is a unique minimizer function $\mathcal{U}$ for $\mathcal{F}(\cdot, \mathbb{R}^p)$. Two striking phenomena concerning the sets $\tilde{O}_J$ are described next:

1. For every point $\hat{u} \in \mathbb{R}^p$, there is a polyhedron $Q_{\hat{u}} \subset \mathbb{R}^p$ of dimension $\#\mathcal{J}(\hat{u})$, such that for every $v \in Q_{\hat{u}}$, the same point $\mathcal{U}(v) = \hat{u}$ is the unique minimizer of $\mathcal{F}(\cdot, v)$.
2. For every $J \subset \{1, \ldots, p-1\}$, there is a subset $\tilde{O}_J \subset \mathbb{R}^p$, composed of $2^{p-\#J-1}$ unbounded polyhedra (of dimension p) of $\mathbb{R}^p$ such that for every $v \in \tilde{O}_J$, the

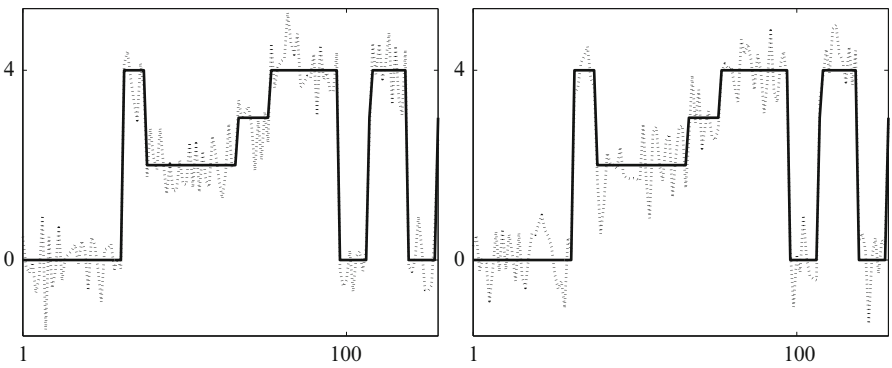


Fig. 9 Data $v = u_o + n$ (—) corresponding to the original u_o (-.-.) contaminated with two different noise samples n on the *left* and on the *right*

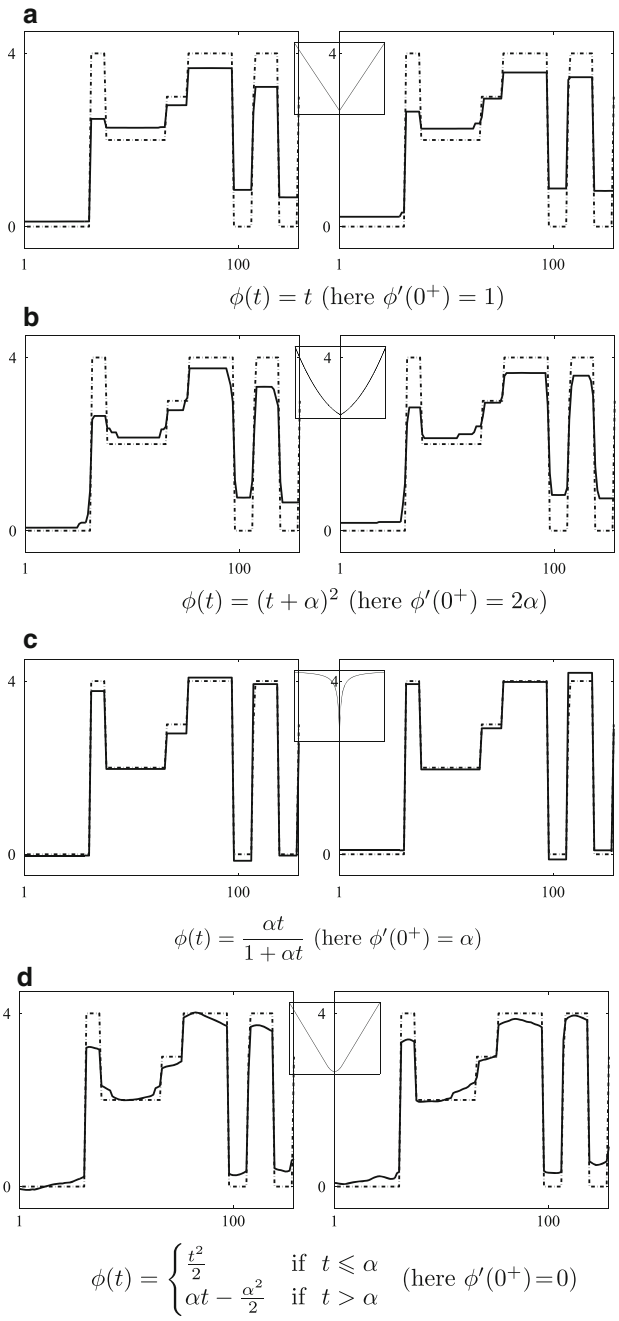


Fig. 10 Restoration using different functions ϕ. Original u_o (-.-.), minimizer $\hat{u}$ (—). Each figure from (**a**) to (**d**) shows the two minimizers $\hat{u}$ corresponding to the two data sets in Fig. 9 (left and right), while the shape of ϕ is plotted in the middle

minimizer $\hat{u}$ of $\mathcal{F}(\cdot, v)$ satisfies $\hat{u}_i = \hat{u}_{i+1}$ for all $i \in J$ and $\hat{u}_i \neq \hat{u}_{i+1}$ for all $i \in J^c$. *Their closure forms a covering of* $\mathbb{R}^p$.

The next Remark 2 deserves to be combined with the conclusions of section "Nonasymptotic Bounds on Minimizers."

Remark 2. The energy in (27) has a straightforward Bayesian interpretation in terms of *maximum a posteriori* (MAP) estimation (see section "Background," first item). The quadratic data-fidelity term corresponds to an observation model of the form $v = Au_o + n$ where n is independent identically distributed (i.i.d.) Gaussian noise with mean zero and variance denoted by σ^2. The likelihood reads $\pi(v|u) = \exp\left(-\frac{1}{2\sigma^2}\|Au - v\|_2^2\right)$. The regularization term corresponds to an i.i.d. Laplacian prior on each difference $u[i]-u[i+1]$, $1 \leqslant i \leqslant p-1$, that is, $\exp(-\lambda|t|)$ for $\lambda = \frac{\beta}{2\sigma^2}$. Since this density is continuous on $\mathbb{R}$, the probability to get a null sample, $t = u[i]-u[i+1] = 0$, is equal to *zero*. However, the results presented above show that for the minimizer $\hat{u}$ of $\mathcal{F}(\cdot, v)$, the probability to have $\hat{u}[i]-\hat{u}[i+1] = 0$ for a certain amount of indexes i is *strictly positive*. This means that the Laplacian prior on the differences $u[i] - u[i+1]$ is far from being incorporated in the MAP solution $\hat{u}$.

Applications

The use of nondifferentiable (and also nonconvex) regularization in compressive sensing is actually extremely abundant; readers can check, e.g., the textbook [50].

Image reconstruction is computed tomography. The concentration of an isotope in a part of the body provides an image characterizing metabolic functions and local blood flow [21, 62]. In emission computed tomography (ECT), a radioactive drug is introduced in a region of the body and the emitted photons are recorded around it. Data are formed by the number of photons $v[i] \geqslant 0$ reaching each detector, $i = 1, \ldots, q$. The observed photon counts v have a Poissonian distribution [21, 90]. Their mean is determined using projection operators $\{a_i,\ i = 1, 2, \ldots, q\}$ and a constant $\rho > 0$. The data-fidelity Ψ derived from the log-likelihood function is nonstrictly convex and reads:

$$\Psi(u, v) = \rho\left\langle \sum_{i=1}^{q} a_i,\ u \right\rangle - \sum_{i=1}^{q} v[i] \ln\left(\langle a_i, u\rangle\right). \tag{28}$$

Figure 11 presents image reconstruction from simulated ECT data by minimizing and energy of the forms (21) and (22) where Ψ is given by (28) and $\{\mathrm{D}_i\}$ yield the first-order differences between each pixel and its eight nearest neighbors. One observes, yet again, that a PF ϕ which is nonconvex with $\phi'(0^+) > 0$ leads to a nicely segmented piecewise constant reconstruction.

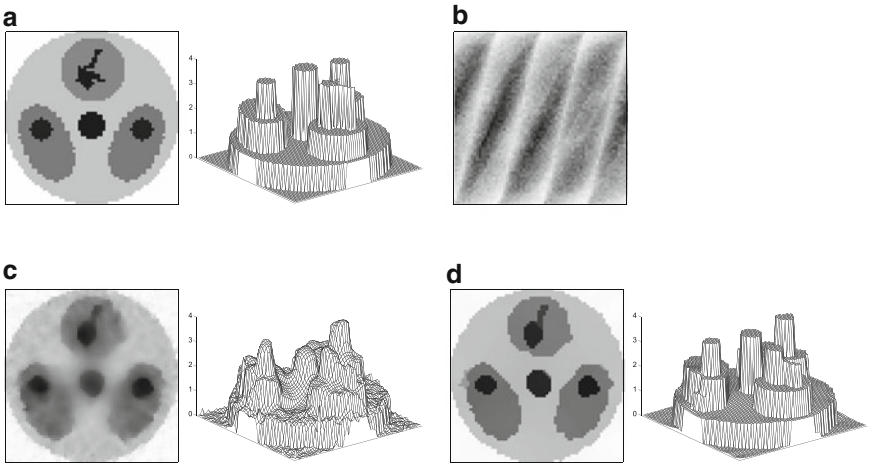


Fig. 11 ECT. $\mathcal{F}(u,v) = \Psi(u,v) + \beta \sum_{i\in I} \phi(|\mathrm{D}_i u|)$. (**a**) Original phantom. (**b**) ECT simulated data. (**c**) $\phi'(0) = 0$, edge preserving. (**d**) $\phi(t) = t/(\alpha + t)$ ($\phi'(0^+) > 0$, nonconvex)

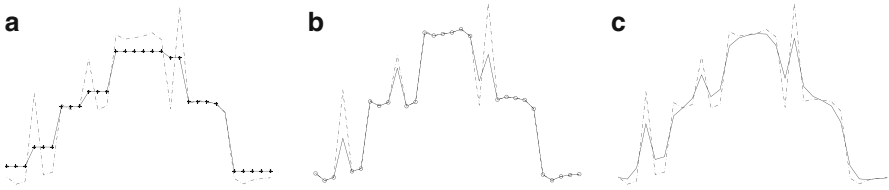


Fig. 12 D is a first-order difference operator, i.e., $\mathrm{D}_i u = u[i] - u[i+1]$, $1 \leqslant i \leqslant p-1$. Data (- - -), restored signal (—). Constant pieces in (**a**) are emphasized using "$*$," while data samples that are equal to the relevant samples of the minimizer in (**b**) are emphasized using "$\circ$"

6 Nonsmooth Data Fidelity

► Figure 12 shows that there is a striking distinction in the behavior of the minimizers relevant to nonsmooth data-fidelity terms (b) with respect to nonsmooth regularization (a). More precisely, many data samples are *fitted exactly* when the data-fidelity term is nonsmooth. This particular behavior is explained and generalized in the present section.

Consider

$$\mathcal{F}(u,v) = \Psi(u,v) + \beta\Phi(u), \tag{29}$$

$$\Psi(u,v) = \sum_{i=1}^{q} \psi\left(|\langle a_i, u\rangle - v[i]|\right), \tag{30}$$

where $a_i \in \mathbb{R}^p$ for all $i \in \{1, \ldots, q\}$ and $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ is a function satisfying

H12 *ψ is $\mathcal{C}^m$, $m \geqslant 2$ on $\mathbb{R}_+^*$ and $\psi'(0^+) > 0$ is finite.*

By this condition, $t \mapsto \psi(|t|)$ is continuous on $\mathbb{R}$. Let $A \in \mathbb{R}^{q \times p}$ denote the matrix such that for any $i = 1, \ldots, q$, its ith row reads a_i^*.

Nonsmooth data-fidelity terms Ψ in energies of the form (29) and (30) were introduced in *image processing* in 2001 [77].

General Results

Here we present some results on the minimizers $\hat{u}$ of $\mathcal{F}$ as given in (29) and (30), where Ψ *is nondifferentiable*, obtained in [78,79]. An additional assumption is that

H13 *The regularization term $\Phi : \mathbb{R}^p \to \mathbb{R}$ in (29) is $\mathcal{C}^m$, $m \geqslant 2$.*

Note that Φ in (29) *can be convex or nonconvex*. To analyze the observation in Fig. 12b, the following set-valued function $\mathcal{J}$ will be useful:

$$(u, v) \in (\mathbb{R}^p \times \mathbb{R}^q) \;\; \mapsto \;\; \mathcal{J}(u, v) = \Big\{ i \in \{1, \ldots, q\} : \langle a_i, u \rangle = v[i] \Big\}. \tag{31}$$

Given v and a (local) minimizer $\hat{u}$ of $\mathcal{F}(\cdot, v)$, the set of all data entries $v[i]$ that are fitted exactly by $(A\hat{u})[i]$ reads $\hat{J} = \mathcal{J}(\hat{u}, v)$. Its complement is $\hat{J}^c = \{1, \ldots, q\} \setminus \hat{J}$.

Theorem 11. *Let $\mathcal{F}$ be of the form (29)–(30) where assumptions H12 and H13 hold. Given $v \in \mathbb{R}^q$, let $\hat{u} \in \mathbb{R}^p$ be a (local) minimizer of $\mathcal{F}(\cdot, v)$. For $\hat{J} = \mathcal{J}(\hat{u}, y)$, where $\mathcal{J}$ is defined according to (31), let*

$$\mathcal{K}_{\hat{J}}(v) = \{u \in \mathbb{R}^p : \; \langle a_i, u \rangle = v[i] \; \forall i \in \hat{J} \text{ and } \langle a_i, u \rangle \neq v[i] \; \forall i \in \hat{J}^c\},$$

and let $K_{\hat{J}}$ be its tangent. Suppose the following:

1. *The set $\Big\{ a_i : i \in \hat{J} \Big\}$ is linearly independent.*
2. *$\forall w \in K_{\hat{J}} \setminus \{0\}$ we have $D_1(\mathcal{F}|_{\overline{\mathcal{K}_{\hat{J}}(v)}})(\hat{u}, v)w = 0$ and $D_1^2(\mathcal{F}|_{\overline{\mathcal{K}_{\hat{J}}(v)}})(\hat{u}, v)(w, w) > 0$.*
3. *$\forall w \in K_{\hat{J}}^{\perp} \setminus \{0\}$ we have $\delta_1 \mathcal{F}(\hat{u}, v)(w) > 0$.*

Then there are a neighborhood $O_{\hat{J}} \subset \mathbb{R}^q$ containing v and a $\mathcal{C}^{m-1}$ local minimizer function $\mathcal{U} : O_{\hat{J}} \to \mathbb{R}^p$ relevant to $\mathcal{F}(\cdot, O_{\hat{J}})$, yielding in particular $\hat{u} = \mathcal{U}(v)$, and

$$v \in O_{\hat{J}} \;\; \Rightarrow \;\; \begin{cases} \langle a_i, \mathcal{U}(v) \rangle = v[i] \text{ if } i \in \hat{J}, \\ \langle a_i, \mathcal{U}(v) \rangle \neq v[i] \text{ if } i \in \hat{J}^c. \end{cases} \tag{32}$$

The result in (32) means that $\mathcal{J}(\mathcal{U}(v), v) = \hat{J}$ is constant on $O_{\hat{J}}$.

Note that for every v and $J \neq \varnothing$, the set $\mathcal{K}_J(v)$ is a finite union of connected components, whereas its closure $\overline{\mathcal{K}_J(v)}$ is an affine subspace. Its tangent $K_{\hat{J}}$ reads

$$K_{\hat{J}} = \{u \in \mathbb{R}^p : \langle a_i, u\rangle = 0 \; \forall i \in \hat{J}\}.$$

A comparison with $K_{\hat{J}}$ in (24) may be instructive. Compare also (b) and (c) in Theorem 11 with (a) and (b) in Theorem 10. By the way, conditions (b) and (c) in Theorem 11 ensure that $\mathcal{F}(\cdot, v)$ reaches a strict minimum at $\hat{u}$ [78, Proposition 1]. Observe that this sufficient condition for strict minimum involves the behavior of $\mathcal{F}(\cdot, v)$ on two orthogonal subspaces separately. This occurs because of the nonsmoothness of $t \mapsto \psi(|t|)$ at zero. It can be useful to note that at a minimizer $\hat{u}$,

$$\begin{aligned}\delta_1\mathcal{F}(\hat{u}, v)(w) = \phi'(0^+)\sum_{i\in\hat{J}} |\langle a_i, w\rangle| + \sum_{i\in\hat{J}^c} \psi'(\langle a_i, \hat{u}\rangle - v[i])\langle a_i, w\rangle \\ +\beta D\Phi(\hat{u})w \geqslant 0, \text{for any } w \in \mathbb{R}^p \end{aligned} \tag{33}$$

Commentary on the assumptions. Assumption (a) does not require the independence of the whole set $\{a_i : i \in \{1, \ldots, q\}\}$. It is easy to check that this assumption fails to hold only for some v is included in a subspace of dimension strictly smaller than q. Hence, assumption (a) is satisfied for almost all $v \in \mathbb{R}^q$ and the theorem addresses *any matrix* A, whether it be singular or invertible.

Assumption (b) is the classical sufficient condition for a strict local minimum of a smooth function over an affine subspace; see Lemma 1 (section "Reminders and Definitions"). If an arbitrary function $\mathcal{F}(\cdot, v) : \mathbb{R}^p \to \mathbb{R}$ has a minimum at $\hat{u}$, then necessarily $\delta_1\mathcal{F}(\hat{u}, v)(w) \geqslant 0$ for all $w \in K_{\hat{J}}^{\perp}$; see Theorem 1. In comparison, (c) requires only that the latter inequality be strict.

It will be interesting to characterize the sets of data v for which (b) and (c) may fail at some (local) minimizers. Some ideas from section "Local Minimizers" can provide a starting point.

Corollary 2. *Let $\mathcal{F}$ be of the form (29)–(30) where $p = q$, and H12 and H13 hold true. Given $v \in \mathbb{R}^q$, let $\hat{u} \in \mathbb{R}^p$ be a (local) minimizer of $\mathcal{F}(\cdot, v)$. Suppose the following:*

(a) *The set $\{a_i : 1 \leqslant i \leqslant q\}$ is linearly independent.*

(b) *$\forall w \in \mathbb{R}^q$ satisfying $\|w\|_2 = 1$ we have $\beta\,|D\Phi(\hat{u})w| < \psi'(0^+)\sum_{i=1}^{q} |\langle a_i, w\rangle|$.*

Then

$$\hat{J} = \{1, \ldots, q\}$$

and there are a neighborhood $O_{\hat{J}} \subset \mathbb{R}^q$ containing v and a $\mathcal{C}^{m-1}$ local minimizer function $\mathcal{U} : O_{\hat{J}} \to \mathbb{R}^p$ relevant to $\mathcal{F}(\cdot, O_{\hat{J}})$, yielding in particular $\hat{u} = \mathcal{U}(v)$, and

$$v \in O_{\hat{J}} \;\Rightarrow\; \langle a_i, \mathcal{U}(v)\rangle = v[i] \quad \forall i \in \hat{J} = \{1, \dots, q\}. \tag{34}$$

More precisely, $\mathcal{U}(v) = A^{-1}v$ *for any* $v \in O_{\hat{J}}$.

In the context of Corollary 2, A is invertible. Combining this with (33) and (b) shows that

$$\mathcal{K}_{\hat{J}}(v) = \{u \in \mathbb{R}^p : \; Au = v\} = A^{-1}v,$$
$$K_{\hat{J}} = \ker(A) = \{0\}.$$

Then

$$\left\{ v \in \mathbb{R}^q : \beta \left| D\Phi(A^{-1}v)w \right| < \psi'(0^+) \sum_{i=1}^{q} |\langle a_i, w\rangle|, \; \forall w \in \mathbb{R}^q \setminus \{0\}, \|w\|_2 = 1 \right\}$$
$$\subset O_{\hat{J}} \equiv O_{\{1,\dots,q\}}.$$

The subset on the left contains an open subset of $\mathbb{R}^q$ by the continuity of $v \mapsto D\Phi(A^{-1}v)$ combined with (b).

Significance of the results. Consider that $\#J \geqslant 1$. The result in (32) means that the set-valued function $v \to \mathcal{J}(\mathcal{U}(v), v)$ is constant on $O_{\hat{J}}$, i.e., that $\mathcal{J}$ is constant under small perturbations of v. Equivalently, all residuals $\langle a_i, \mathcal{U}(v)\rangle - v[i]$ for $i \in \hat{J}$ are null on $O_{\hat{J}}$.

Theorem 11 shows that $\mathbb{R}^q$ contains *volumes of positive measure* composed of data that lead to local minimizers which fit exactly the data entries belonging to the same set. In general, there are volumes corresponding to various $\hat{J}$ so that noisy data come across them. That is why *nonsmooth data-fidelity terms generically yield minimizers fitting exactly a certain number of the data entries*. The resultant numerical effect is observed in Fig. 12b as well as in Figs. 14 and 15.

Remark 3 (Stability of Minimizers). The fact that there is a $\mathcal{C}^{m-1}$ local minimizer function shows that, in spite of the nonsmoothness of $\mathcal{F}$, for any v, all local minimizers of $\mathcal{F}(\cdot, v)$ which satisfy the conditions of the theorem are *stable under weak perturbations of data* v. This result extends Lemma 1.

Example 6. Let $\mathcal{F}$ read

$$\mathcal{F}(u, v) = \sum_{i=1}^{q} |u[i] - v[i]| + \frac{\beta}{2} \sum_{i=1}^{q} (u[i])^2, \quad \beta > 0.$$

It is easy to see that there is a unique local minimizer function $\mathcal{U}$ which is given by

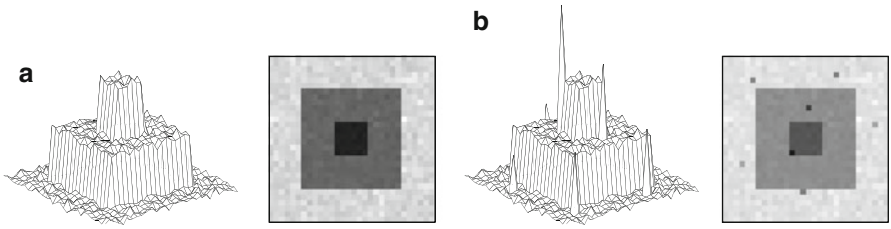


Fig. 13 Original u_o and data v degraded by outliers. (**a**) Original u_o. (**b**) Data $v = u \cdot \text{outliers}$

$$\mathcal{U}(v)[i] = \frac{1}{\beta}\text{sign}(v[i]) \quad \text{if} \quad |v[i]| > \frac{1}{\beta},$$
$$\mathcal{U}(v)[i] = v[i] \quad \text{if} \quad |v[i]| \leqslant \frac{1}{\beta}.$$

Condition (c) in Theorem 11 fails to hold only for $\left\{v \in \mathbb{R}^q : |v[i]| = \frac{1}{\beta},\ \forall i \in \hat{J}\right\}$. This set is of Lebesgue measure zero in $\mathbb{R}^q$. For any $J \in \{1, \ldots, q\}$ put

$$O_J = \left\{ v \in \mathbb{R}^q : |v[i]| \leqslant \frac{1}{\beta},\ \forall i \in J \quad \text{and} \quad |v[i]| > \frac{1}{\beta},\ \forall i \in J^c \right\}.$$

Obviously, every $v \in O_J$ gives rise to a minimizer $\hat{u}$ satisfying

$$\hat{u}[i] = v[i], \quad \forall i \in J \ \text{ and } \ \hat{u}[i] \neq v[i], \quad \forall i \in J^c.$$

Each set O_J has a positive Lebesgue measure in $\mathbb{R}^q$. Moreover, *the union of all O_J when J ranges on all subsets $J \subset \{1, \ldots, q\}$ (including the empty set) forms a partition of $\mathbb{R}^q$.*

Numerical experiment. The original image u_o is shown in Fig. 13a. Data v in Fig. 13b are obtained by replacing some pixels of u_o by aberrant impulsions, called *outliers*.

In all Figs. 14–17, $\{\mathrm{D}_i\}$ correspond to the first-order differences between each pixel and its four nearest neighbors. Figure 14a corresponds to an ℓ_1 data-fidelity term for $\beta = 0.14$. The outliers are well visible although their amplitudes are clearly reduced. The image of the residuals $v - \hat{u}$, shown in Fig. 14b, is null everywhere except at the positions of the outliers in v. The pixels corresponding to nonzero residuals (i.e., the elements of $\hat{J}^c$) provide a good estimate of the locations of the outliers in v. Figure 15a shows a minimizer $\hat{u}$ of the same $\mathcal{F}(\cdot, v)$ obtained for $\beta = 0.25$. This minimizer does not contain visible outliers and is very close to the original image u_o. The image of the residuals $v - \hat{u}$ in Fig. 15b is null only on restricted areas but has a very small magnitude everywhere beyond the outliers.

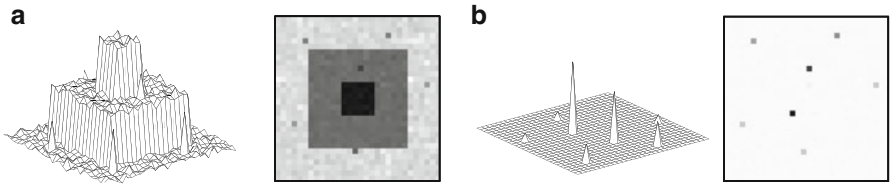


Fig. 14 Restoration using $\mathcal{F}(u,v) = \sum_i |u[i] - v[i]| + \beta \sum_{i\in I} |\mathrm{D}_i u|^\alpha$ $\alpha = 1.1$ and $\beta = 0.14$. (**a**) Restoration $\hat{u}$ for $\beta = 0.14$. (**b**) Residual $v - \hat{u}$

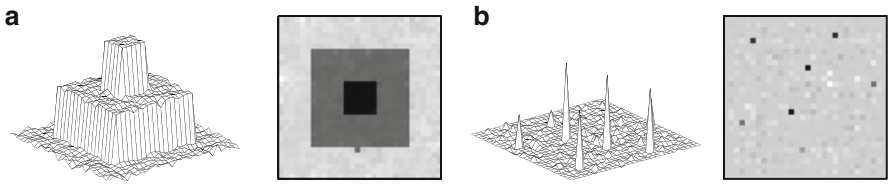


Fig. 15 Restoration using $\mathcal{F}(u,v) = \sum_i |u[i]-v[i]|+\beta \sum_{i\in I} |\mathrm{D}_i u|^\alpha$ for $\alpha = 1.1$ and $\beta = 0.25$. (**a**) Restoration $\hat{u}$ for $\beta = 0.25$. (**b**) Residual $v - \hat{u}$

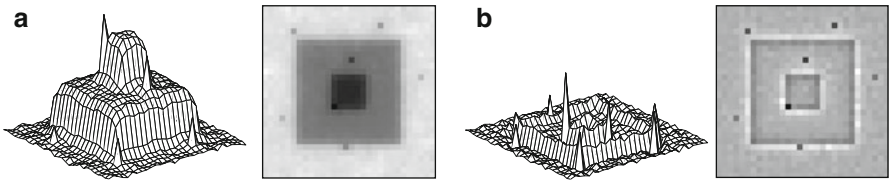


Fig. 16 Restoration using a smooth energy, $\mathcal{F}(u,v) = \sum_i (u[i]-v[i])^2 + \beta \sum_i (|\mathrm{D}_i u|)^2$, $\beta = 0.2$. (**a**) Restoration $\hat{u}$ for $\beta = 0.2$. (**b**) Residual $v - \hat{u}$

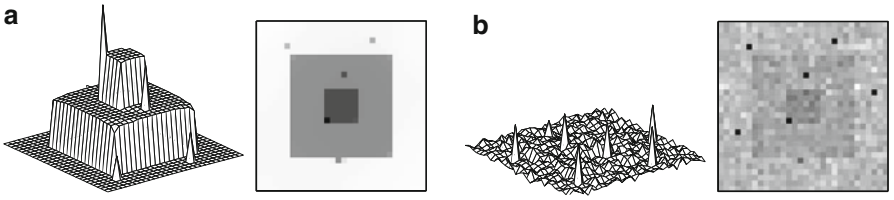


Fig. 17 Restoration using nonsmooth regularization $\mathcal{F}(u,v) = \sum_i (u[i] - v[i])^2 + \beta \sum_i |\mathrm{D}_i u|$, $\beta = 0.2$. (**a**) Restoration $\hat{u}$ for $\beta = 0.2$. (**b**) Residual $v - \hat{u}$

The minimizers of two different cost-functions $\mathcal{F}$ involving a *smooth* data-fidelity term Ψ, shown in Figs. 16 and 17, do not fit any data entry. In particular, the restoration in Fig. 17 corresponds to a nonsmooth regularization and it is constant over large regions; this effect was explained in Sect. 5.

Applications

The possibility to keep some data samples unchanged by using nonsmooth data fidelity is a precious property in various application fields. Nonsmooth data fidelities are good to detect and smooth outliers. This property was exploited for deblurring under impulse noise contamination; see, e.g., [10–12].

Denoising of frame coefficients. Consider the recovery of an original (unknown) $u_o \in \mathbb{R}^p$ – a signal or an image containing smooth zones and edges – from noisy data

$$v = u_o + n,$$

where n represents a perturbation. As discussed in Sect. 4, a systematic default of the images and signals restored using *convex* edge-preserving PFs ϕ is that the amplitude of edges is underestimated.

Shrinkage estimators operate on a decomposition of data v into a frame of ℓ^2, say $\{w_i : i \in J\}$ where J is a set of indexes. Let W be the corresponding frame operator, i.e., $(Wv)[i] = \langle v, w_i \rangle$, $\forall i \in J$, and $\tilde{W}$ be a left inverse of W, giving rise to the dual frame $\{\tilde{w}_i : i \in J\}$. The frame coefficients of v read $y = Wv$ and are contaminated with noise Wn. The inaugural work of Donoho and Johnstone [40] considers two different shrinkage estimators: given $T > 0$, *hard thresholding* corresponds to

$$y_T[i] = \begin{cases} y[i] \text{ if } i \in J_1, \\ 0 \quad \text{if } i \in J_0, \end{cases} \quad \text{where} \quad \begin{cases} J_0 = \{i \in J : |y[i]| \leqslant T\}; \\ J_1 = J \backslash J_0, \end{cases} \tag{35}$$

while in soft thresholding one takes $y_T[i] = y[i] - T\,\mathrm{sign}(y[i])$ if $i \in J_1$ and $y_T[i] = 0$ if $i \in J_0$. Both soft and hard thresholding are asymptotically optimal in the minimax sense if n is white Gaussian noise of standard deviation σ and

$$T = \sigma \sqrt{2 \log_e p}. \tag{36}$$

This threshold is difficult to use in practice because it increases with the size of u. Numerous improvements were realized; see, e.g., [4, 13, 24, 34, 38, 66, 70]. In all cases, the main problem is that smoothing large coefficients oversmooths edges, while thresholding small coefficients can generate Gibbs-like oscillations near edges; see Fig. 18c, d. If shrinkage is weak, noisy coefficients (outliers) remain almost unchanged and produce artifacts having the shape of $\{\tilde{w}_i\}$; see Fig. 18c–e.

In order to alleviate these difficulties, several authors proposed *hybrid methods* where the information contained in important coefficients $y[i]$ is combined with priors in the domain of the sought-after signal or image [18,25,36,43,67]. A critical analysis was presented in [46].

A specialized hybrid method involving ℓ_1 data fidelity on frame coefficients is proposed in [46]. Data are initially hard thresholded – see (35) – using a *suboptimal*

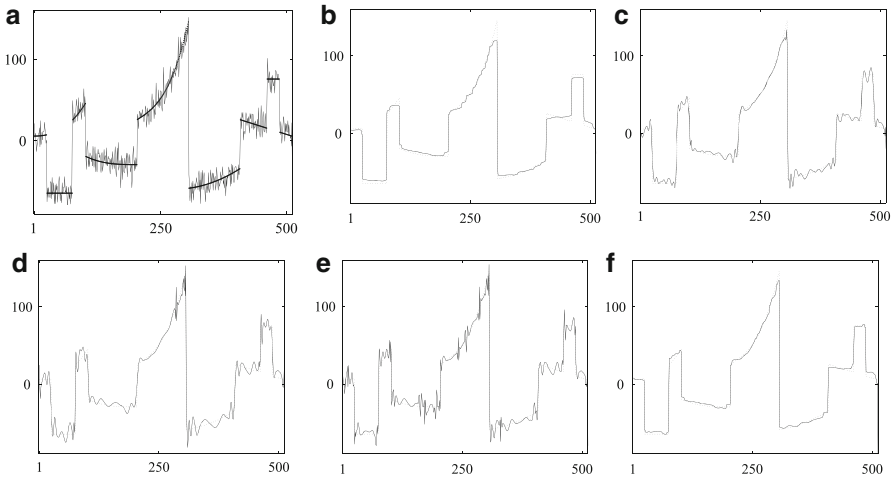


Fig. 18 Methods to restore the noisy signal in (**a**). Restored signal (—), original signal (- -). (**a**) Original and data corrupted with white Gaussian noise. (**b**) TV regularization. (**c**) *Sure-shrink*. (**d**) $T = 35$ optimal, $\hat{u}_T = \sum_i y_T[i]\tilde{w}_i$. (**e**) y_T, $T = 23$, $\hat{u}_T = \sum_i y_T[i]\tilde{w}_i$. (**f**) The proposed method

threshold T in order to keep as much as possible information. (The use of another shrinkage estimator would alter all coefficients, which is not desired.) Then

1. J_1 is composed of:

   - Large coefficients bearing the main features of u_o that one wishes to preserve intact
   - Aberrant coefficients (outliers) that must be restored using the regularization term

2. J_0 is composed of:

   - Noise coefficients that must be kept null.
   - Coefficients $y[i]$ corresponding to edges and other details in u_o – these need to be restored in accordance with the prior incorporated in the regularization term.

In order to reach the goals formulated in 1 and 2 above, denoised coefficients $\hat{x}$ are defined as a minimizer of the hybrid energy $F(., y)$ given below:

$$F(x, y) = \lambda_1 \sum_{i \in J_1} |x[i] - y[i]| + \lambda_0 \sum_{i \in J_0} |x[i]| + \sum_{i \in I} \phi\left(\|\mathrm{D}_i \tilde{W} x\|_2\right), \quad \lambda_{0,1} > 0, \tag{37}$$

where ϕ is convex and edge preserving. Then the sought-after denoised image or signal is

$$\hat{u} = \tilde{W}\hat{x} = \sum_{i \in J} \tilde{w}_i \hat{x}[i].$$

Several properties relevant to the minimizers of F in (37), the parameters λ_i, $i \in \{0, 1\}$, and the solution $\hat{u}$ are outlined in [46].

Noisy data v are shown along with the original u_o in Fig. 18a. The restoration in Fig. 18b minimizes $\mathcal{F}(u) = \|Au - v\|_2^2 + \beta\text{TV}$ – homogeneous regions remain noisy, edges are smoothed, and spikes are eroded. Figure 18c is obtained using the *sure-shrink* method [41] from the toolbox WaveLab. The other restorations use thresholded Daubechies wavelet coefficients with eight vanishing moments. The optimal value for the hard thresholding obtained using (36) is $T = 35$. The relevant restoration – Fig. 18d – exhibits important Gibbs-like oscillations as well as wavelet-shaped artifacts. For $T = 23$ the coefficients have a richer information content, but $\tilde{W}y_T$, shown in Fig. 18e, manifests Gibbs artifacts and many wavelet-shaped artifacts. Introducing the thresholded coefficients of Fig. 18e in the specialized energy F in (37) leads to Fig. 18f: edges are clean and piecewise polynomial parts are well recovered.

7 Nonsmooth Data Fidelity and Regularization

The L_1-TV Case

For discrete signals of finite length, energies of the form $\mathcal{F}(u, v) = \|u - v\|_1 + \beta \sum_{i=1}^{p-1} |u[i+1] - u[i]|$ were considered by Alliney in 1992 [1].

Following [1, 78, 79], S. Esedoglu and T. Chan explored in [28] the minimizers of the L_1-TV functional given below

$$\mathcal{F}(u, v) = \int_{\mathbb{R}^d} |u(x) - v(x)| dx + \beta \int_{\mathbb{R}^d} |\nabla u(x)| dx, \tag{38}$$

where the sought-after minimizer $\hat{u}$ belongs to the space of bounded variation functions on $\mathbb{R}^d$. The main focus is on images, i.e., $d = 2$. The analysis in [28] is based on a representation of $\mathcal{F}$ in (38) in terms of the level sets of u and v. Most of the results are established for data v given by the characteristic function χ_Σ of a bounded domain $\Sigma \subset \mathbb{R}^d$. Theorem 5.2 in [28] says that if $v = \chi_\Sigma$, where $\Sigma \subset \mathbb{R}^d$ is bounded, then $\mathcal{F}(\cdot, v)$ admits a minimizer of the form $\hat{u} = \chi_{\hat{\Sigma}}$ (with possibly $\hat{\Sigma} \neq \Sigma$). Furthermore, Corollary 5.3. in [28] states that if in addition Σ is *convex*, then for almost every $\beta \geqslant 0$, $\mathcal{F}(\cdot, v)$ admits a unique minimizer and $\hat{u} = \chi_{\hat{\Sigma}}$ with $\hat{\Sigma} \subseteq \Sigma$. Moreover, it is shown that small features in the image maintain their contrast intact up to some value of β, while for a larger β they suddenly disappear.

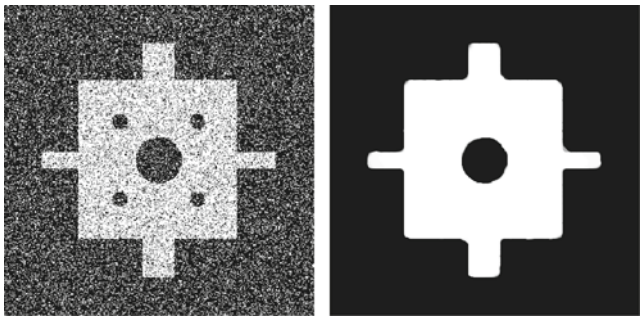


Fig. 19 Restoration of a binary noisy image by minimizing L_1-TV

Denoising of Binary Images and Convex Relaxation

Many problems such as text denoising and document processing, two-phase image segmentation, shape restoration, and fairing of surfaces in computer graphics are naturally stated as the minimization of an energy over the set of the binary images. These energies are obviously nonconvex since the constraint set is finite. Their global minimizer was shown in [29] to be also the minimizer of the convex L_1-TV functional which is convex. This result yielded much simpler algorithms for binary image restoration. An illustration is given on Fig. 19.

Since then, L_1-TV relaxations have became a common tool for convex relaxations; see, e.g., among many others [84] and the references therein.

Also, L_1-TV energies were revealed very successful in image decomposition; see, e.g., [8, 48].

Multiplicative Noise Removal

In various active imaging systems, such as synthetic aperture radar (SAR), laser, or ultrasound imaging, the data representing the underlying (unknown image) S_0 are corrupted with multiplicative noise. Such a noise is a severe degradation; see Fig. 20. When possible, a few independent measurements for the same scene are realized, $\Sigma_k = S_0\eta_k$ for $k \in \{1, \ldots, K\}$, where the noise η_k is typically modeled by the one-sided exponential distribution. The data Σ used for denoising is the average of the set of all K measurements:

$$\Sigma = \frac{1}{K}\sum_{k=1}^{K}\Sigma_k = S_0\eta. \tag{39}$$

The combined multiplicative noise follows a Gamma distribution. Adaptive filtering works only if the noise is weak. For strong noise, variational methods often use TV

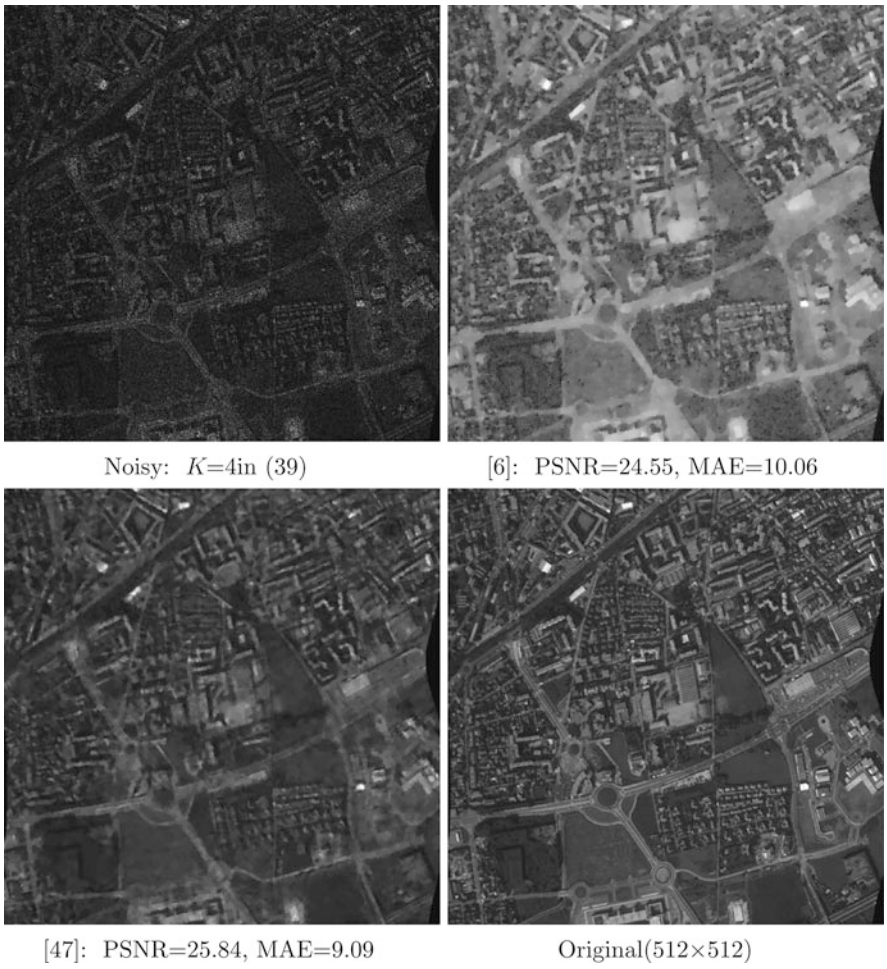


Fig. 20 Aerial image of the town of Nîmes (512 × 512) for $K = 4$ in (39). Restorations using different methods. Parameters: [6] for $\rho = 120$; [47] $T = 2\sqrt{\psi_1(K)}$, $\lambda_0 = 1.5, \lambda_1 = 10$

regularization. In [6] the log-likelihood of the raw data (39) is regularized using TV. Instead, the properties of L_1-TV are used to design an energy in [47]. First, the log-data $v = \log \Sigma$ is decomposed into a curvelet transform yielding noisy coefficients $y = Wv$. A suboptimal hard thresholding is applied for T adapted to the expectation of the log noise. Let $I_0 = \{i \ : \ |y[i]| \leqslant T\}$ and $I_1 = \{i \ : \ |y[i]| > T\}$. Since the threshold is low, I_1 contains outliers. Coefficients $\hat{x}$ are restored by minimizing

$$F(x) = \lambda_1 \sum_{i \in I_1} |(x - y)[i]| + \lambda_0 \sum_{i \in I_0} |x[i]| + \mathrm{TV}(x) \ .$$

The restored image $\hat{S}$, shown in Fig. 20, is obtained as $\hat{S} = \exp(\tilde{W}(\hat{x}))\ B$ where $\tilde{W}$ is a left inverse of W and B is a bias correction.

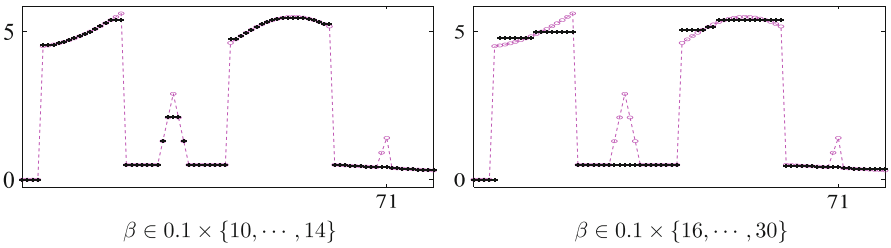


Fig. 21 Minimizers of $\mathcal{F}(\cdot, v)$ as given in (40) for $\phi(t) = \ln(\alpha t + 1)$, $\alpha = 2$ and different values of β. Data samples (o o o), minimizer samples $\hat{u}[i]$ (+ + +)

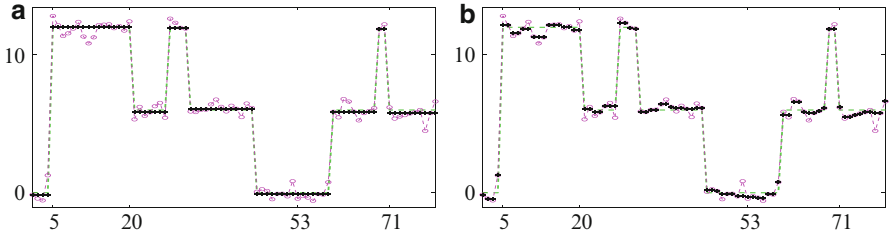


Fig. 22 Minimizers of $\mathcal{F}(\cdot, v)$ as given in (40) for different PFs ϕ. Data are corrupted with Gaussian noise. Data samples $v[i]$ are marked with (o o o), samples $\hat{u}[i]$ of the minimizer – with (+ + +). The original signal is reminded in (– – –). (**a**) $\phi(t) = \frac{\alpha t}{\alpha t+1}$, $\alpha = 4$, $\beta = 3$. (**b**) $\phi(t) = t$, $\beta = 0.8$

ℓ_1 Data Fidelity with Regularization Concave on $\mathbb{R}_+$

One could expect that ℓ_1 data fidelity regularized with a PF concave on $\mathbb{R}_+$ should somehow reinforce the properties of $\ell_1 - \mathrm{TV}$. The question was recently examined in [83]. Consider the energy

$$\mathcal{F}(u, v) = \sum_{i \in I} |\langle a_i, u \rangle - v[i]| + \beta \sum_{j \in J} \phi(|\mathrm{D}_i u|) \tag{40}$$

$$\text{for} \quad I \stackrel{\text{def}}{=} \{1, \ldots, q\} \quad \text{and} \quad J \stackrel{\text{def}}{=} \{1, \ldots, r\}$$

where $\phi : \mathbb{R}_+ \to \mathbb{R}_+$ is continuous and concave on $\mathbb{R}_+$ (e.g., (f10), (f11), and (f12) in Table 1).

Motivation

Figures 21 and 22 depict (the global) minimizers of $\mathcal{F}(u, v)$ in (40) for a one-dimensional signal where $A = \mathrm{Id}$, $\{\mathrm{D}_i\}$ are first-order differences, and ϕ is smooth and concave on $\mathbb{R}_+$.

The tests in Figs. 21 and 22 show that a PF concave on $\mathbb{R}_+$ considerably reinforces the properties of $\ell_1 - \mathrm{TV}$. One observes that the minimizer satisfies *exactly*

part of the data term and part of the prior term (corresponding to constant pieces). In Fig. 22b, the previous $\ell_1 - \mathrm{TV}$ model is considered. Figure 21 shows also that the minimizer remains unchanged for some range of values of β and that after a threshold value, it is simplified.

Example 7 below furnishes a first intuition on the reasons underlying the phenomena observed in Figs. 21 and 22.

Example 7. Given $v \in \mathbb{R}$, consider the function $\mathcal{F}(\cdot\,, v) : \mathbb{R} \to \mathbb{R}$ given below

$$\mathcal{F}(u, v) = |u - v| + \beta\phi(|u|) \ \text{ for } \phi \text{ obeying H} \,. \tag{41}$$

The necessary conditions for $\mathcal{F}(\cdot\,, v)$ to have a (local) minimum at $\hat{u} \neq 0$ and $\hat{u} \neq v$ – that its first differential meets $D_1\mathcal{F}(\hat{u}, v) = 0$ and that its second differential obeys $D_1^2\mathcal{F}(\hat{u}, v) \geqslant 0$ – do not hold:

$$\hat{u} \notin \{0, v\} \quad \Rightarrow \quad \begin{cases} D_1\mathcal{F}(\hat{u}, v) = \operatorname{sign}(\hat{u} - v) + \beta\varphi'(|\hat{u}|)\operatorname{sign}(\hat{u}) = 0 \,, \\ \\ D_1^2\mathcal{F}(\hat{u}, v) = \beta\varphi''(|\hat{u}|) < 0 \,, \end{cases}$$

where the last inequality comes from the strict concavity of ϕ on $\mathbb{R}_+^*$. *Hence, $\mathcal{F}(\cdot\,, v)$ cannot have a minimizer such that $\hat{u} \neq 0$ and $\hat{u} \neq v$, for any $v \in \mathbb{R}$.* Being coercive, $\mathcal{F}(\cdot\,, v)$ does have minimizers. Consequently, any (local) minimizer of $\mathcal{F}$ in (41) satisfies

$$\hat{u} \in \{0, v\} \,.$$

Main Theoretical Results

The PFs considered here are concave on $\mathbb{R}_+$ and smooth on $\mathbb{R}_+^*$. More precisely, they satisfy H1 (Sect. 1), H8, and H10 (section "Assumptions on Potential Functions ϕ"). One can see Fig. 3, right, for an illustration of the assumptions.

Proposition 3. *Let $\mathcal{F}(\cdot\,, v)$ read as in (40). Assume that H3 (Sect. 3) holds and that ϕ satisfies H1 (Sect. 1), H8, and H10. Then for any v, $\mathcal{F}(\cdot\,, v)$ has global minimizers.*

Given $v \in \mathbb{R}^q$, with *each* $\hat{u} \in \mathbb{R}^p$ the following subsets are associated:

$$\begin{aligned} \hat{I}_0 &\stackrel{\text{def}}{=} \{i \in I \ \mid \ \langle a_i \hat{u} \rangle = v[i]\} \quad \text{and} \quad \hat{I}_0^c \stackrel{\text{def}}{=} I \setminus \hat{I}_0, \\ \hat{J}_0 &\stackrel{\text{def}}{=} \{i \in J \ \mid \ \mathrm{D}_i \hat{u} = 0\} \qquad \text{and} \quad \hat{J}_0^c \stackrel{\text{def}}{=} J \setminus \hat{J}_0 \,. \end{aligned} \tag{42}$$

Proposition 4. *For $\mathcal{F}(\cdot\,, v)$ as in (40) satisfying H1, H8, and H10, let $\hat{u}$ be a (local) minimizer of $\mathcal{F}(\cdot\,, v)$. Then*

$$\big(\hat{I}_0 \cup \hat{J}_0\big) \neq \varnothing \,.$$

H14 *The point* $\hat{u} \in \mathbb{R}^p$ *is such that* $\hat{I}_0 \neq \varnothing$ *and that*

$$w \in \ker \mathrm{D} \setminus \{0\} \quad \Rightarrow \quad \exists i \in \hat{I}_0 \ \text{ such that } \ \langle a_i, w \rangle \neq 0 \,. \tag{43}$$

If $\operatorname{rank} \mathrm{D} = p$, then (43) is trivial. Anyway, (43) is not a strong requirement.

Theorem 12. *Consider* $\mathcal{F}(\cdot, v)$*, as given in* (40)*, satisfying H3, as well as H1, H8, and H10. Let* $\hat{u}$ *be a (local) minimizer of* $\mathcal{F}(\cdot, v)$ *meeting* $\hat{J}_0^c \neq \varnothing$ *and H14. Then* $\hat{u}$ *is the* unique *solution of the* full column rank linear system *given below*

$$\begin{cases} \langle a_i, w \rangle = v[i] & \forall i \in \hat{I}_0 \,, \\ \mathrm{D}_j w = 0 & \forall j \in \hat{J}_0 \,. \end{cases} \tag{44}$$

Significance of the Results

An immediate consequence of Theorem 12 is the following:

► Each (local) minimizer of $\mathcal{F}(\cdot, v)$ is strict.

Another consequence is that the matrix H with rows $\left(a_i^*, \forall i \in \hat{I}_0 \text{ and } \mathrm{D}_j, \forall j \in \hat{J}_0\right)$ has *full column rank*. This provides a strong *necessary condition for a (local) minimizer* of $\mathcal{F}(\cdot, v)$. And since $\hat{u}$ in (44) solves a linear system, it involves the same kind of "contrast invariance" as the $L_1 - \mathrm{TV}$ model. A detailed inspection of the minimizers in Figs. 21 and 22 corroborate Theorem 12. A more practical interpretation of this result reads as follows:

► Each pixel of a (local) minimizer $\hat{u}$ of $\mathcal{F}(\cdot, v)$ is involved in (at least) one data equation that is fitted exactly $\langle a_i, \hat{u} \rangle = v[i]$ or in (at least) one vanishing operator $\mathrm{D}_j \hat{u} = 0$ or in both types of equations.

This remarkable property can be used in different ways.

Applications

An energy $\mathcal{F}(\cdot, v)$ of the form in (40) with a PF ϕ strictly concave on $\mathbb{R}_+$ is a good choice when

- There are some nearly faithful data points $v[i]$;
- The matrix D provides a very reliable prior on the sought-after solution.

A natural way for such a prior is to construct for D an application-dependent dictionary.

MR Image Reconstruction from Highly Undersampled Data

In the experiment in Fig. 23, only **5 %** *randomly* chosen noisy data samples in the k-space (i.e., individual noisy Fourier coefficients) are available; see (a). Data are

Fig. 23 Reconstructed images from **5 %** noisy randomly selected samples in the k-space using different methods. (**a**) Zero-filling Fourier recovery. (**b**) $\ell_2 - \mathrm{TV}$. (**c**) $\mathcal{F}(\cdot, v)$ in (40)

contaminated with $\mathrm{SNR} = 37\,\mathrm{dB}$ white centered *Gaussian noise*. This is a highly underdetermined, ill-posed inverse problem. It can be related to compressed sensing in MRI; see, e.g., [58]. The Shepp-Logan phantom being locally constant with oval shapes, the linear operators $\{\mathrm{D}_i\}$ in (40) yield the usual discrete gradient of the image, so that the regularization term provides a correct prior. Indeed, $\mathrm{D}u_o$ is the sparsest linear transform for this image. Clearly, A is the undersampled Fourier transform corresponding to the 5 % randomly chosen k-samples. For Gaussian noise, an ℓ_2 quadratic data fitting term is a classical choice. The $\ell_2 - \mathrm{TV}$ cost-function $\|Au - v\|_2^2 + \beta \mathrm{TV}(u)$ is the standard tool to solve this kind of problems. The result is shown in Fig. 23b.

8 Conclusion

This chapter provided some theoretical results relating the shape of the energy $\mathcal{F}$ to minimize and the salient features of its minimizers $\hat{u}$ (see (7), section "The Main Features of the Minimizers as a Function of the Energy"). These results can serve as a kind of *inverse modeling*: given an inverse problem along with our requirements (priors) on its solution, they guide us how to construct an energy functional whose minimizers properly incorporate all this information. The theoretical results are illustrated using numerical examples. Various application fields can take a benefit from these results. The problem of such an inverse modeling remains open because of the diversity of the inverse problems to solve and the possible energy functionals.

Cross-References

- ▸ Compressive Sensing
- ▸ Inverse Scattering
- ▸ Iterative Solution Methods
- ▸ Linear Inverse Problems

- Mathematical Methods in PET and SPECT Imaging
- Mumford and Shah Model and Its Applications to Image Segmentation and Image Restoration
- Regularization Methods for Ill-Posed Problems
- Statistical Methods in Imaging
- Total Variation in Imaging

References

1. Alliney, S.: Digital filters as absolute norm regularizers. IEEE Trans. Signal Process. **SP-40**, 1548–1562 (1992)
2. Alter, F., Durand, S., Forment, J.: Adapted total variation for artifact free decompression of JPEG images. J. Math. Imaging Vis. **23**, 199–211 (2005)
3. Ambrosio, L., Fusco, N., Pallara, D.: Functions of Bounded Variation and Free Discontinuity Problems. Oxford Mathematical Monographs. Oxford University Press (2000)
4. Antoniadis, A., Fan, J.: Regularization of wavelet approximations. J. Acoust. Soc. Am. **96**, 939–967 (2001)
5. Attouch, H., Bolte, J., Svaiter, B.F.: Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized Gauss–Seidel methods. Math. Program. **137**, 91–129 (2013)
6. Aubert, G., Aujol, J.-F.: A variational approach to remove multiplicative noise. SIAM J. Appl. Math. **68**, 925–946 (2008)
7. Aubert, G., Kornprobst, P.: Mathematical Problems in Image Processing, 2nd edn. Springer, Berlin (2006)
8. Aujol, J.-F., Gilboa, G., Chan, T., Osher, S.: Structure-texture image decomposition – modeling, algorithms, and parameter selection. Int. J. Comput. Vis. **67**, 111–136 (2006)
9. Auslender, A., Teboulle, M.: Asymptotic Cones and Functions in Optimization and Variational Inequalities. Springer, New York (2003)
10. Bar, L., Brook, A., Sochen, N., Kiryati, N.: Deblurring of color images corrupted by impulsive noise. IEEE Trans. Image Process. **16**, 1101–1111 (2007)
11. Bar, L., Kiryati, N., Sochen, N.: Image deblurring in the presence of impulsive noise. Int. J. Comput. Vis. **70**, 279–298 (2006)
12. Bar, L., Sochen, N., Kiryati, N.: Image deblurring in the presence of salt-and-pepper noise. In: Proceeding of 5th International Conference on Scale Space and PDE Methods in Computer Vision, Hofgeismar. LNCS, vol. 3459, pp. 107–118 (2005)
13. Belge, M., Kilmer, M., Miller, E.: Wavelet domain image restoration with adaptive edge-preserving regularization. IEEE Trans. Image Process. **9**, 597–608 (2000)
14. Besag, J.E.: Spatial interaction and the statistical analysis of lattice systems (with discussion). J. R. Stat. Soc. B **36**, 192–236 (1974)
15. Besag, J.E.: Digital image processing: towards Bayesian image analysis. J. Appl. Stat. **16**, 395–407 (1989)
16. Black, M., Rangarajan, A.: On the unification of line processes, outlier rejection, and robust statistics with applications to early vision. Int. J. Comput. Vis. **19**, 57–91 (1996)
17. Blake, A., Zisserman, A.: Visual Reconstruction. MIT, Cambridge (1987)
18. Bobichon, Y., Bijaoui, A.: Regularized multiresolution methods for astronomical image enhancement. Exp. Astron. **7**, 239–255 (1997)
19. Bolte, J., Sabach, S., Teboulle, M.: Proximal alternating linearized minimization for nonconvex and nonsmooth problems. Math. Program. Ser. A **146**(1-2), 459–494 (2014)
20. Bouman, C., Sauer, K.: A generalized Gaussian image model for edge-preserving map estimation. IEEE Trans. Image Process. **2**, 296–310 (1993)

21. Bouman, C., Sauer, K.: A unified approach to statistical tomography using coordinate descent optimization. IEEE Trans. Image Process. **5**, 480–492 (1996)
22. Bredies, K., Holler, M.: Regularization of linear inverse problems with total generalized variation. J. Inverse Ill-Posed Probl. (2014). doi:10.1515/jip-2013-0068
23. Bredies, K., Kunich, K., Pock, T.: Total generalized variation. SIAM J. Imaging Sci. **3**(3), 480–491 (2010)
24. Candès, E.J., Donoho, D., Ying, L.: Fast discrete curvelet transforms. SIAM Multiscale Model. Simul. **5**, 861–899 (2005)
25. Candès, E.J., Guo, F.: New multiscale transforms, minimum total variation synthesis. Applications to edge-preserving image reconstruction. Signal Process. **82**, 1519–1543 (2002)
26. Catte, F., Coll, T., Lions, P.L., Morel, J.M.: Image selective smoothing and edge detection by nonlinear diffusion (I). SIAM J. Numer. Anal. **29**, 182–193
27. Chambolle, A.: An algorithm for total variation minimization and application. J. Math. Imaging Vis. **20**, 89–98 (2004)
28. Chan, T., Esedoglu, S.: Aspects of total variation regularized L^1 function approximation. SIAM J. Appl. Math. **65**, 1817–1837 (2005)
29. Chan, T., Esedoglu, S., Nikolova, M.: Algorithms for finding global minimizers of image segmentation and denoising models. SIAM J. Appl. Math. **66**, 1632–1648 (2006)
30. Chan, T.F., Wong, C.K.: Total variation blind deconvolution. IEEE Trans. Image Process. **7**, 370–375 (1998)
31. Charbonnier, P., Blanc-Féraud, L., Aubert, G., Barlaud, M.: Deterministic edge-preserving regularization in computed imaging. IEEE Trans. Image Process. **6**, 298–311 (1997)
32. Chellapa, R., Jain, A.: Markov random fields: theory and application. Academic, Boston (1993)
33. Chen, X., Ng, M., Zhang, C.: Non-Lipschitz ℓ_p-regularization and box constrained model for image restoration. IEEE Trans. Image Process. **21**, 4709–4721 (2012)
34. Chesneau, C., Fadili, J., Starck, J.-L.: Stein block thresholding for image denoising. Appl. Comput. Harmon. Anal. **28**, 67–88 (2010)
35. Ciarlet, P.G.: Introduction to Numerical Linear Algebra and Optimization. Cambridge University Press, Cambridge (1989)
36. Coifman, R.R., Sowa, A.: Combining the calculus of variations and wavelets for image enhancement. Appl. Comput. Harmon. Anal. **9**, 1–18 (2000)
37. Demoment, G.: Image reconstruction and restoration: overview of common estimation structure and problems. IEEE Trans. Acoust. Speech Signal Process. **ASSP-37**, 2024–2036 (1989)
38. Do, M.N., Vetterli, M.: The contourlet transform: an efficient directional multiresolution image representation. IEEE Trans. Image Process. **15**, 1916–1933 (2005)
39. Dobson, D., Santosa, F.: Recovery of blocky images from noisy and blurred data. SIAM J. Appl. Math. **56**, 1181–1199 (1996)
40. Donoho, D.L., Johnstone, I.M.: Ideal spatial adaptation by wavelet shrinkage. Biometrika **81**, 425–455 (1994)
41. Donoho, D.L., Johnstone, I.M.: Adapting to unknown smoothness via wavelet shrinkage. J. Acoust. Soc. Am. **90**, 1200–1224 (1995)
42. Dontchev, A.L., Zollezi, T.: Well-Posed Optimization Problems. Springer, New York (1993)
43. Durand, S., Froment, J.: Reconstruction of wavelet coefficients using total variation minimization. SIAM J. Sci. Comput. **24**, 1754–1767 (2003)
44. Durand, S., Nikolova, M.: Stability of minimizers of regularized least squares objective functions I: study of the local behavior. Appl. Math. Optim. (Springer, New York) **53**, 185–208 (2006)
45. Durand, S., Nikolova, M.: Stability of minimizers of regularized least squares objective functions II: study of the global behaviour. Appl. Math. Optim. (Springer, New York) **53**, 259–277 (2006)
46. Durand, S., Nikolova, M.: Denoising of frame coefficients using ℓ^1 data-fidelity term and edge-preserving regularization. SIAM J. Multiscale Model. Simul. **6**, 547–576 (2007)
47. Durand, S., Fadili, J., Nikolova, M.: Multiplicative noise removal using L1 fidelity on frame coefficients. J. Math. Imaging Vis. **36**, 201–226 (2010)

48. Duval, V., Aujol, J.-F., Gousseau, Y.: The TVL1 model: a geometric point of view. SIAM J. Multiscale Model. Simul. **8**, 154–189 (2009)
49. Ekeland, I., Temam, R.: Convex analysis and variational problems. North-Holland/SIAM, Amsterdam (1976)
50. Eldar, Y.C., Kutyniok, G.: Compressed Sensing: Theory and Applications. Cambridge University Press (1212)
51. Fessler, F.: Mean and variance of implicitly defined biased estimators (such as penalized maximum likelihood): applications to tomography. IEEE Trans. Image Process. **5**, 493–506 (1996)
52. Fiacco, A., McCormic, G.: Nonlinear Programming. Classics in Applied Mathematics. SIAM, Philadelphia (1990)
53. Geman, D.: Random fields and inverse problems in imaging. In: École d'Été de Probabilités de Saint-Flour XVIII – 1988. Lecture Notes in Mathematics, vol. 1427, pp. 117–193. Springer, Berlin/Heidelberg (1990)
54. Geman, D., Reynolds, G.: Constrained restoration and recovery of discontinuities. IEEE Trans. Pattern Anal. Mach. Intell. **PAMI-14**, 367–383 (1992)
55. Geman, D., Yang, C.: Nonlinear image recovery with half-quadratic regularization. IEEE Trans. Image Process. **IP-4**, 932–946 (1995)
56. Geman, S., Geman, D.: Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Trans. Pattern Anal. Mach. Intell. **PAMI-6**, 721–741 (1984)
57. Green, P.J.: Bayesian reconstructions from emission tomography data using a modified EM algorithm. IEEE Trans. Med. Imaging **MI-9**, 84–93 (1990)
58. Lustig, M., Donoho, D., Santos, J.M., Pauly, L.M.: Compressed sensing MRI: a look how CS can improve our current imaging techniques: IEEE Signal Proc. Mag. **25**, 72–82 (2008)
59. Haddad, A., Meyer, Y.: Variational methods in image processing. In: Perspective in Nonlinear Partial Differential equations in Honor of Haïm Brezis. Contemporary Mathematics, vol. 446, pp. 273–295. AMS, Providence (2007)
60. Hiriart-Urruty, J.-B., Lemaréchal, C.: Convex Analysis and Minimization Algorithms, vols. I, II. Springer, Berlin (1996)
61. Hofmann, B.: Regularization for applied inverse and ill posed problems. Teubner, Leipzig (1986)
62. Kak, A., Slaney, M.: Principles of Computerized Tomographic Imaging. IEEE, New York (1987)
63. Keren, D., Werman, M.: Probabilistic analysis of regularization. IEEE Trans. Pattern Anal. Mach. Intell. **PAMI-15**, 982–995 (1993)
64. Li, S.: Markov Random Field Modeling in Computer Vision, 1st edn. Springer, New York (1995)
65. Li, S.Z.: On discontinuity-adaptive smoothness priors in computer vision. IEEE Trans. Pattern Anal. Mach. Intell. **PAMI-17**, 576–586 (1995)
66. Luisier, F., Blu, T.: SURE-LET multichannel image denoising: interscale orthonormal wavelet thresholding. IEEE Trans. Image Process. **17**, 482–492 (2008)
67. Malgouyres, F.: Minimizing the total variation under a general convex constraint for image restoration. IEEE Trans. Image Process. **11**, 1450–1456 (2002)
68. Morel, J.-M., Solimini, S.: Variational Methods in Image Segmentation. Birkhäuser, Basel (1995)
69. Morozov, V.A.: Regularization Methods for Ill Posed Problems. CRC, Boca Raton (1993)
70. Moulin, P., Liu, J.: Analysis of multiresolution image denoising schemes using generalized Gaussian and complexity priors. IEEE Trans. Image Process. **45**, 909–919 (1999)
71. Moulin, P., Liu, J.: Statistical imaging and complexity regularization. IEEE Trans. Inf. Theory **46**, 1762–1777 (2000)
72. Mumford, D., Shah, J.: Optimal approximations by piecewise smooth functions and associated variational problems. Commun. Pure Appl. Math. **42**, 577–684 (1989)
73. Nashed, M., Scherzer, O.: Least squares and bounded variation regularization with nondifferentiable functional. Numer. Funct. Anal. Optim. **19**, 873–901 (1998)

74. Nikolova, M.: Regularisation functions and estimators. In: Proceedings of the IEEE International Conference on Image Processing, Lausanne, vol. 2, pp. 457–460 (1996)
75. Nikolova, M.: Estimées localement fortement homogènes. Comptes-Rendus de l'Acad émie des Sciences **325**(série 1), 665–670 (1997)
76. Nikolova, M.: Thresholding implied by truncated quadratic regularization. IEEE Trans. Image Process. **48**, 3437–3450 (2000)
77. Nikolova, M.: Image restoration by minimizing objective functions with nonsmooth data-fidelity terms. In: IEEE International Conference on Computer Vision/Workshop on Variational and Level-Set Methods, Vancouver, pp. 11–18 (2001)
78. Nikolova, M.: Minimizers of cost-functions involving nonsmooth data-fidelity terms. Application to the processing of outliers. SIAM J. Numer. Anal. **40**, 965–994 (2002)
79. Nikolova, M.: A variational approach to remove outliers and impulse noise. J. Math. Imaging Vis. **20**, 99–120 (2004)
80. Nikolova, M.: Weakly constrained minimization. Application to the estimation of images and signals involving constant regions. J. Math. Imaging Vis. **21**, 155–175 (2004)
81. Nikolova, M.: Analysis of the recovery of edges in images and signals by minimizing nonconvex regularized least-squares. SIAM J. Multiscale Model. Simul. **4**, 960–991 (2005)
82. Nikolova, M.: Analytical bounds on the minimizers of (nonconvex) regularized least-squares. AIMS J. Inverse Probl. Imaging **1**, 661–677 (2007)
83. Nikolova, M., Ng, M., Tam, C.P.: On ℓ_1 data fitting and concave regularization for image recovery. SIAM J. Sci. Comput. **35**, 397–430 (2013)
84. Papadakis, N., Yildizoglu, R., Aujol, J.-F., Caselles, V.: High-dimension multi-label problems: convex or non convex relaxation? SIAM J. Imaging Sci. **6**, 2603–2639 (2013)
85. Perona, P., Malik, J.: Scale-space and edge detection using anisotropic diffusion. IEEE Trans. Pattern Anal. Mach. Intell. **PAMI-12**, 629–639 (1990)
86. Rockafellar, R.T., Wets, J.B.: Variational Analysis. Springer, New York (1997)
87. Rudin, L., Osher, S., Fatemi, C.: Nonlinear total variation based noise removal algorithm. Physica D **60**, 259–268 (1992)
88. Robini, M., Magnin, I.: Optimization by stochastic continuation. SIAM J. Imaging Sci. **3**, 1096–1121 (2010)
89. Robini, M., Reissman, P.-J.: From simulated annealing to stochastic continuation: a new trend in combinatorial optimization. J. Glob. Optim. **56**, 185–215 (2013)
90. Sauer, K., Bouman, C.: A local update strategy for iterative reconstruction from projections. IEEE Trans. Signal Process. **SP-41**, 534–548 (1993)
91. Scherzer, O., Grasmair, M., Grossauer, H., Haltmeier, M., Lenzen, F.: Variational Problems in Imaging. Springer, New York (2009)
92. Tautenhahn, U.: Error estimates for regularized solutions of non-linear ill posed problems. Inverse Probl. **10**, 485–500 (1994)
93. Tikhonov, A., Arsenin, V.: Solutions of Ill Posed Problems. Winston, Washington, D.C. (1977)
94. Vogel, C.: Computational Methods for Inverse Problems. Frontiers in Applied Mathematics Series, vol. 23. SIAM, New York (2002)
95. Welk, M., Steidl, G., Weickert, J.: Locally analytic schemes: a link between diffusion filtering and wavelet shrinkage. Appl. Comput. Harmon. Anal. **24**, 195–224 (2008)
96. Winkler, G.: Image Analysis, Random Fields and Markov Chain Monte Carlo Methods. A Mathematical Introduction. Applications of Mathematics. Stochastic Models and Applied Probability, vol. 27, 2nd edn. Springer, Berlin (2006)

Compressive Sensing

Massimo Fornasier and Holger Rauhut

Contents

1 Introduction 206
2 Background 208
Early Developments in Applications 209
Sparse Approximation 210
Information-Based Complexity and Gelfand Widths 210
Compressive Sensing 211
Developments in Computer Science 211
3 Mathematical Modelling and Analysis 212
Preliminaries and Notation 212
Sparsity and Compression 213
Compressive Sensing 214
The Null Space Property 216
The Restricted Isometry Property 217
Coherence 220
RIP for Gaussian and Bernoulli Random Matrices 221
Random Partial Fourier Matrices 222
Compressive Sensing and Gelfand Widths 224
Extensions of Compressive Sensing 227
Applications 232
4 Numerical Methods 233
A Primal-Dual Algorithm 233
Iteratively Re-weighted Least Squares 236
Numerical Experiments 243
Extensions to Affine Low-Rank Minimization 245
5 Open Questions 247

M. Fornasier (✉)
Faculty of Mathematics, Technische Universität München, Garching, Germany
e-mail: massimo.fornasier@ma.tum.de

H. Rauhut
Lehrstuhl C für Mathematik, RWTH Aachen University, Aachen, Germany
e-mail: rauhut@mathc.rwth-aachen.de

© Springer Science+Business Media New York 2015

O. Scherzer (ed.), *Handbook of Mathematical Methods in Imaging*,
DOI 10.1007/978-1-4939-0790-8_6

Deterministic Compressed Sensing Matrices 247
Removing Log-Factors in the Fourier-RIP Estimate 248
Compressive Sensing with Nonlinear Measurements 248
6 Conclusion 248
Cross-References 249
Recommended Reading 249
References 249

Abstract

Compressive sensing is a recent type of sampling theory, which predicts that sparse signals and images can be reconstructed from what was previously believed to be incomplete information. As a main feature, efficient algorithms such as ℓ_1-minimization can be used for recovery. The theory has many potential applications in signal processing and imaging. This chapter gives an introduction and overview on both theoretical and numerical aspects of compressive sensing.

1 Introduction

The traditional approach of reconstructing signals or images from measured data follows the well-known Shannon sampling theorem [155], which states that the sampling rate must be twice the highest frequency. Similarly, the fundamental theorem of linear algebra suggests that the number of collected samples (measurements) of a discrete finite-dimensional signal should be at least as large as its length (its dimension) in order to ensure reconstruction. This principle underlies most devices of current technology, such as analog to digital conversion, medical imaging, or audio and video electronics. The novel theory of compressive sensing (CS) – also known under the terminology of compressed sensing, compressive sampling, or sparse recovery – provides a fundamentally new approach to data acquisition which overcomes this common wisdom. It predicts that certain signals or images can be recovered from what was previously believed to be highly incomplete measurements (information). This chapter gives an introduction to this new field. Both fundamental theoretical and algorithmic aspects are presented, with the awareness that it is impossible to retrace in a few pages all the current developments of this field, which was growing very rapidly in the past few years and undergoes significant advances on an almost daily basis.

CS relies on the empirical observation that many types of signals or images can be well approximated by a sparse expansion in terms of a suitable basis, that is, by only a small number of nonzero coefficients. This is the key to the efficiency of many lossy compression techniques such as JPEG, MP3, etc. A compression is obtained by simply storing only the largest basis coefficients. When reconstructing the signal, the non-stored coefficients are simply set to zero. This is certainly a reasonable strategy when full information of the signal is available. However, when the signal first has to be acquired by a somewhat costly, lengthy, or otherwise difficult

measurement (sensing) procedure, this seems to be a waste of resources: First, large efforts are spent in order to obtain full information on the signal, and afterwards most of the information is thrown away at the compression stage. One might ask whether there is a clever way of obtaining the compressed version of the signal more directly, by taking only a small number of measurements of the signal. It is not obvious at all whether this is possible since measuring directly the large coefficients requires to know a priori their location. Quite surprisingly, compressive sensing provides nevertheless a way of reconstructing a compressed version of the original signal by taking only a small amount of *linear* and *nonadaptive* measurements. The precise number of required measurements is comparable to the compressed size of the signal. Clearly, the measurements have to be suitably designed. It is a remarkable fact that all provably good measurement matrices designed so far are random matrices. It is for this reason that the theory of compressive sensing uses a lot of tools from probability theory.

It is another important feature of compressive sensing that practical reconstruction can be performed by using efficient algorithms. Since the interest is in the vastly undersampled case, the linear system describing the measurements is underdetermined and therefore has infinitely many solutions. The key idea is that the sparsity helps in isolating the original vector. The first naive approach to a reconstruction algorithm consists in searching for the sparsest vector that is consistent with the linear measurements. This leads to the combinatorial ℓ_0-problem (see (4) below), which unfortunately is NP-hard in general. There are essentially two approaches for tractable alternative algorithms. The first is convex relaxation leading to ℓ_1-minimization – also known as basis pursuit (see (5)) – while the second constructs greedy algorithms. This overview focuses on ℓ_1-minimization. By now basic properties of the measurement matrix which ensure sparse recovery by ℓ_1-minimization are known: the *null space property (NSP)* and the *restricted isometry property (RIP)*. The latter requires that all column submatrices of a certain size of the measurement matrix are well conditioned. This is where probabilistic methods come into play because it is quite hard to analyze these properties for deterministic matrices with minimal amount of measurements. Among the provably good measurement matrices are Gaussian, Bernoulli random matrices, and partial random Fourier matrices.

Figure 1 serves as a first illustration of the power of compressive sensing. It shows an example for recovery of a 10-sparse signal $x \in \mathbb{C}^{300}$ from only 30 samples (indicated by the red dots in Fig. 1b). From a first look at the time-domain signal, one would rather believe that reconstruction should be impossible from only 30 samples. Indeed, the spectrum reconstructed by traditional ℓ_2-minimization is very different from the true spectrum. Quite surprisingly, ℓ_1-minimization performs nevertheless an exact reconstruction, that is, with no recovery error at all!

An example from nuclear magnetic resonance imaging serves as a second illustration. Here, the device scans a patient by taking 2D or 3D frequency measurements within a radial geometry. Figure 2a describes such a sampling set of a 2D Fourier transform. Since a lengthy scanning procedure is very uncomfortable for the patient,

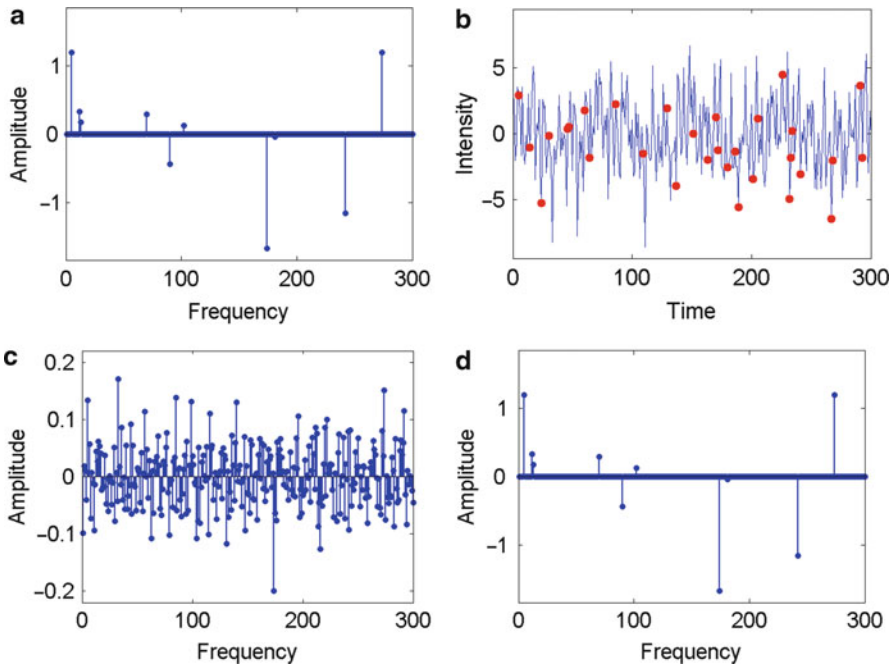


Fig. 1 (**a**) 10-sparse Fourier spectrum, (**b**) time-domain signal of length 300 with 30 samples, (**c**) reconstruction via ℓ_2-minimization, (**d**) exact reconstruction via ℓ_1-minimization

it is desired to take only a minimal amount of measurements. Total variation minimization, which is closely related to ℓ_1-minimization, is then considered as recovery method. For comparison, Fig. 2b shows the recovery by a traditional back-projection algorithm. Figure 2c, d displays iterations of an algorithm, which was proposed and analyzed in [72] to perform efficient large-scale total variation minimization. The reconstruction in Fig. 2d is again exact!

2 Background

Although the term compressed sensing (compressive sensing) was coined only recently with the paper by Donoho [47], followed by a huge research activity, such a development did not start out of thin air. There were certain roots and predecessors in application areas such as image processing, geophysics, medical imaging, computer science, as well as in pure mathematics. An attempt is made to put such roots and current developments into context below, although only a partial overview can be given due to the numerous and diverse connections and developments.

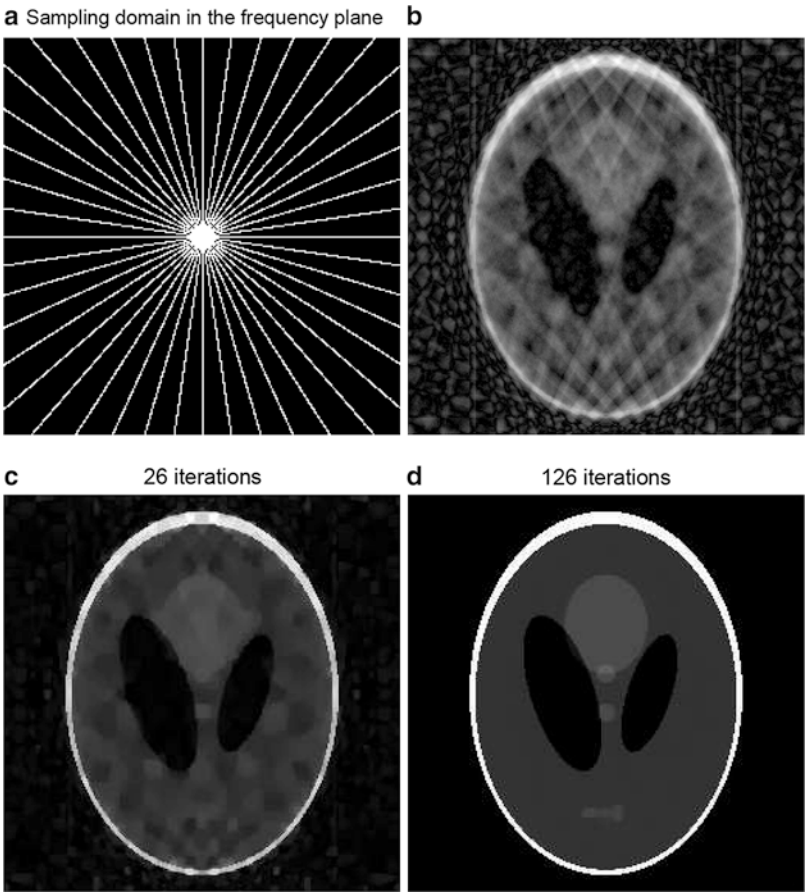


Fig. 2 (**a**) Sampling data of the NMR image in the Fourier domain which corresponds to only 0.11 % of all samples. (**b**) Reconstruction by back projection. (**c**) Intermediate iteration of an efficient algorithm for large-scale total variation minimization. (**d**) The final reconstruction is exact

Early Developments in Applications

Presumably the first algorithm which can be connected to sparse recovery is due to the French mathematician de Prony [127]. The so-called Prony method, which has found numerous applications [109], estimates nonzero amplitudes and corresponding frequencies of a sparse trigonometric polynomial from a small number of equispaced samples by solving an eigenvalue problem. The use of ℓ_1-minimization appears already in the Ph.D. thesis of B. Logan [106] in connection with sparse frequency estimation, where he observed that L_1-minimization may recover exactly a frequency-sparse signal from undersampled data provided the sparsity is small enough. The paper by Donoho and Logan [52] is perhaps the

earliest theoretical work on sparse recovery using L_1-minimization. Nevertheless, geophysicists observed in the late 1970s and 1980s that ℓ_1-minimization can be successfully employed in reflection seismology where a sparse reflection function indicating changes between subsurface layers is sought [140, 148]. In NMR spectroscopy the idea to recover sparse Fourier spectra from undersampled non-equispaced samples was first introduced in the 1990s [158] and has seen a significant development since then.

In image processing the use of total variation minimization, which is closely connected to ℓ_1-minimization and compressive sensing, first appears in the 1990s in the work of Rudin, Osher, and Fatemi [139] and was widely applied later on. In statistics where the corresponding area is usually called *model selection*, the use of ℓ_1-minimization and related methods was greatly popularized with the work of Tibshirani [149] on the so-called LASSO (Least Absolute Shrinkage and Selection Operator).

Sparse Approximation

Many lossy compression techniques such as JPEG, JPEG-2000, MPEG, or MP3 rely on the empirical observation that audio signals and digital images have a sparse representation in terms of a suitable basis. Roughly speaking one compresses the signal by simply keeping only the largest coefficients. In certain scenarios such as audio signal processing, one considers the generalized situation where sparsity appears in terms of a redundant system – a so-called dictionary or frame [36] – rather than a basis. The problem of finding the sparsest representation/approximation in terms of the given dictionary turns out to be significantly harder than in the case of sparsity with respect to a basis where the expansion coefficients are unique. Indeed, in [108, 114], it was shown that the general ℓ_0-problem of finding the sparsest solution of an underdetermined system is NP-hard. Greedy strategies such as matching pursuit algorithms [108], FOCUSS [86] and ℓ_1-minimization [35] were subsequently introduced as tractable alternatives. The theoretical understanding under which conditions greedy methods and ℓ_1-minimization recover the sparsest solutions began to develop with the work in [50, 51, 62, 78, 81, 87, 151, 152].

Information-Based Complexity and Gelfand Widths

Information-based complexity (IBC) considers the general question of how well a function f belonging to a certain class $\mathcal{F}$ can be recovered from n sample values or, more generally, the evaluation of n linear or nonlinear functionals applied to f [150]. The optimal recovery error which is defined as the maximal reconstruction error for the "best" sampling method and "best" recovery method (within a specified class of methods) over all functions in the class $\mathcal{F}$ is closely related to the so-called *Gelfand width* of $\mathcal{F}$ [38, 47, 117]. Of particular interest for compressive sensing is $\mathcal{F} = B_1^N$, the ℓ_1-ball in $\mathbb{R}^N$ since its elements can be well

approximated by sparse ones. A famous result due to Kashin [96] and Gluskin and Garnaev [79, 84] sharply bounds the Gelfand widths of B_1^N (as well as their duals, the *Kolmogorov widths*) from above and below; see also [77]. While the original interest of Kashin was in the estimate of n-widths of Sobolev classes, these results give precise performance bounds in compressive sensing on how well any method may recover (approximately) sparse vectors from linear measurements [38,47]. The upper bounds on Gelfand widths were derived in [96] and [79] using (Bernoulli and Gaussian) random matrices (see also [107]), and in fact such type of matrices have become very useful also in compressive sensing [26, 47].

Compressive Sensing

The numerous developments in compressive sensing began with the seminal work [30,47]. Although key ingredients were already in the air at that time, as mentioned above, the major contribution of these papers was to realize that one can combine the power of ℓ_1-minimization and random matrices in order to show *optimal* results on the ability of ℓ_1-minimization of recovering (approximately) sparse vectors. Moreover, the authors made very clear that such ideas have strong potential for numerous application areas. In their work [26, 30] Candès, Romberg, and Tao introduced the *restricted isometry property* (which they initially called the *uniform uncertainty principle*) which is a key property of compressive sensing matrices. It was shown that Gaussian, Bernoulli, and partial random Fourier matrices [26,129,138] possess this important property. These results require many tools from probability theory and finite-dimensional Banach space geometry, which have been developed for a rather long time now; see, e.g., [95, 103].

Donoho [49] developed a different path and approached the problem of characterizing sparse recovery by ℓ_1-minimization via polytope geometry, more precisely, via the notion of k-neighborliness. In several papers sharp phase transition curves were shown for Gaussian random matrices separating regions where recovery fails or succeeds with high probability [49, 53, 54]. These results build on previous work in pure mathematics by Affentranger and Schneider [2] on randomly projected polytopes.

Developments in Computer Science

In computer science the related area is usually addressed as the *heavy hitters* detection or *sketching*. Here one is interested not only in recovering signals (such as huge data streams on the Internet) from vastly undersampled data, but one requires sublinear runtime in the signal length N of the recovery algorithm. This is no impossibility as one only has to report the locations and values of the nonzero (most significant) coefficients of the sparse vector. Quite remarkably sublinear algorithms are available for sparse Fourier recovery [80]. Such algorithms use ideas from *group testing* which date back to World War II, when Dorfman [56] invented an efficient method for detecting draftees with syphilis.

In sketching algorithms from computer science, one actually designs the matrix and the fast algorithm simultaneously [42, 82]. More recently, *bipartite expander graphs* have been successfully used in order to construct good compressed sensing matrices together with associated fast reconstruction algorithms [11].

3 Mathematical Modelling and Analysis

This section introduces the concept of sparsity and the recovery of sparse vectors from incomplete linear and nonadaptive measurements. In particular, an analysis of ℓ_1-minimization as a recovery method is provided. The *null space property* and the *restricted isometry property* are introduced, and it is shown that they ensure robust sparse recovery. It is actually difficult to show these properties for deterministic matrices and the optimal number m of measurements, and the major breakthrough in compressive sensing results is obtained for random matrices. Examples of several types of random matrices which ensure sparse recovery are given, such as Gaussian, Bernoulli, and partial random Fourier matrices.

Preliminaries and Notation

This exposition mostly treats complex vectors in $\mathbb{C}^N$ although sometimes the considerations will be restricted to the real case $\mathbb{R}^N$. The ℓ_p-norm of a vector $x \in \mathbb{C}^N$ is defined as

$$\|x\|_p := \left(\sum_{j=1}^{N} |x_j|^p \right)^{1/p}, \quad 0 < p < \infty,$$
$$\|x\|_\infty := \max_{j=1,\dots,N} |x_j|. \tag{1}$$

For $1 \le p \le \infty$, it is indeed a norm, while for $0 < p < 1$, it is only a quasi-norm. When emphasizing the norm, the term ℓ_p^N is used instead of $\mathbb{C}^N$ or $\mathbb{R}^N$. The unit ball in ℓ_p^N is $B_p^N = \{x \in \mathbb{C}^N, \|x\|_p \le 1\}$. The operator norm of a matrix $A \in \mathbb{C}^{m \times N}$ from ℓ_p^N to ℓ_p^m is denoted

$$\|A\|_{p \to p} = \max_{\|x\|_p = 1} \|Ax\|_p. \tag{2}$$

In the important special case $p = 2$, the operator norm is the maximal singular value $\sigma_{\max}(A)$ of A.

For a subset $T \subset \{1, \dots, N\}$, one denotes by $x_T \in \mathbb{C}^N$ the vector which coincides with $x \in \mathbb{C}^N$ on the entries in T and is zero outside T. Similarly, A_T denotes the column submatrix of A corresponding to the columns indexed by T. Further, $T^c = \{1, \dots, N\} \setminus T$ denotes the complement of T and $\#T$ or $|T|$ indicates the cardinality of T. The kernel of a matrix A is denoted by $\ker A = \{x, Ax = 0\}$.

Sparsity and Compression

Compressive sensing is based on the empirical observation that many types of real-world signals and images have a sparse expansion in terms of a suitable basis or frame, for instance, a wavelet expansion. This means that the expansion has only a small number of significant terms, or, in other words, that the coefficient vector can be well approximated with one having only a small number of nonvanishing entries.

The support of a vector x is denoted $\operatorname{supp}(x) = \{j : x_j \neq 0\}$ and

$$\|x\|_0 := |\operatorname{supp}(x)|.$$

It has become common to call $\|\cdot\|_0$ the ℓ_0-norm, although it is not even a quasi-norm. A vector x is called *k-sparse* if $\|x\|_0 \leq k$. For $k \in \{1, 2, \ldots, N\}$,

$$\Sigma_k := \{x \in \mathbb{C}^N : \|x\|_0 \leq k\}$$

denotes the set of k-sparse vectors. Furthermore, the *best k-term approximation error* of a vector $x \in \mathbb{C}^N$ in ℓ_p is defined as

$$\sigma_k(x)_p = \inf_{z \in \Sigma_k} \|x - z\|_p.$$

If $\sigma_k(x)$ decays quickly in k, then x is called *compressible*. Indeed, in order to compress x, one may simply store only the k largest entries. When reconstructing x from its compressed version, the non-stored entries are simply set to zero, and the reconstruction error is $\sigma_k(x)_p$. It is emphasized at this point that the procedure of obtaining the compressed version of x is *adaptive* and *nonlinear* since it requires the search of the largest entries of x in absolute value. In particular, the location of the nonzeros is a nonlinear type of information.

The *best k-term approximation* of x can be obtained using the nonincreasing rearrangement $r(x) = (|x_{i_1}|, \ldots, |x_{i_N}|)^T$, where i_j denotes a permutation of the indexes such that $|x_{i_j}| \geq |x_{i_{j+1}}|$ for $j = 1, \ldots, N-1$. Then it is straightforward to check that

$$\sigma_k(x)_p := \left(\sum_{j=k+1}^{N} r_j(x)^p \right)^{1/p}, \quad 0 < p < \infty.$$

And the vector $x_{[k]}$ derived from x by setting to zero all the $N - k$ smallest entries in absolute value is the *best k-term approximation*,

$$x_{[k]} = \arg\min_{z \in \Sigma_k} \|x - z\|_p,$$

for any $0 < p \leq \infty$.

The next lemma states essentially that ℓ_q-balls with small q (ideally $q \leq 1$) are good models for compressible vectors.

Lemma 1. *Let* $0 < q < p \leq \infty$ *and set* $r = \frac{1}{q} - \frac{1}{p}$. *Then*

$$\sigma_k(x)_p \leq k^{-r}, \quad k = 1, 2, \ldots, N \quad \text{for all } x \in B_q^N.$$

Proof. Let T be the set of indexes of the k-largest entries of x in absolute value. The nonincreasing rearrangement satisfies $|r_k(x)| \leq |x_j|$ for all $j \in T$, and therefore

$$k r_k(x)^q \leq \sum_{j \in T} |x_j|^q \leq \|x\|_q^q \leq 1.$$

Hence, $r_k(x) \leq k^{-\frac{1}{q}}$. Therefore,

$$\sigma_k(x)_p^p = \sum_{j \notin T} |x_j|^p \leq \sum_{j \notin T} r_k(x)^{p-q} |x_j|^q \leq k^{-\frac{p-q}{q}} \|x\|_q^q \leq k^{-\frac{p-q}{q}},$$

which implies $\sigma_k(x)_p \leq k^{-r}$. ∎

Compressive Sensing

The above outlined adaptive strategy of compressing a signal x by only keeping its largest coefficients is certainly valid when full information on x is available. If, however, the signal first has to be acquired or measured by a somewhat costly or lengthy procedure, then this seems to be a waste of resources: At first, large efforts are made to acquire the full signal and then most of the information is thrown away when compressing it. One may ask whether it is possible to obtain more directly a compressed version of the signal by taking only a small amount of *linear and nonadaptive* measurements. Since one does not know a priori the large coefficients, this seems a daunting task at first sight. Quite surprisingly, compressive sensing nevertheless predicts that reconstruction from vastly undersampled nonadaptive measurements is possible – even by using efficient recovery algorithms.

Taking m linear measurements of a signal $x \in \mathbb{C}^N$ corresponds to applying a matrix $A \in \mathbb{C}^{m \times N}$ – the *measurement matrix* –

$$y = Ax. \tag{3}$$

The vector $y \in \mathbb{C}^m$ is called the *measurement vector*. The main interest is in the vastly undersampled case $m \ll N$. Without further information, it is, of course, impossible to recover x from y since the linear system (3) is highly underdetermined and has therefore infinitely many solutions. However, if the additional assumption that the vector x is k-sparse is imposed, then the situation dramatically changes as will be outlined.

The approach for a recovery procedure that probably comes first to mind is to search for the sparsest vector x which is consistent with the measurement vector $y = Ax$. This leads to solving the ℓ_0-*minimization problem*

$$\min \|z\|_0 \quad \text{subject to } Az = y. \tag{4}$$

Unfortunately, this combinatorial minimization problem is NP-hard in general [108, 114]. In other words, an algorithm that solves (4) for *any* matrix A and *any* right-hand side y is necessarily computationally intractable. Therefore, essentially two practical and tractable alternatives to (4) have been proposed in the literature: convex relaxation leading to ℓ_1-minimization – also called basis pursuit [35] – and greedy algorithms, such as various matching pursuits [151, 153]. Quite surprisingly for both types of approaches, various recovery results are available, which provide conditions on the matrix A and on the sparsity $\|x\|_0$ such that the recovered solution coincides with the original x and consequently also with the solution of (4). This is no contradiction to the NP-hardness of (4) since these results apply only to a subclass of matrices A and right-hand sides y.

The ℓ_1-minimization approach considers the solution of

$$\min \|z\|_1 \quad \text{subject to } Az = y, \tag{5}$$

which is a convex optimization problem and can be seen as a convex relaxation of (4). Various efficient convex optimization techniques apply for its solution [17]. In the real-valued case, (5) is equivalent to a linear program, and in the complex-valued case, it is equivalent to a second-order cone program. Therefore, standard software applies for its solution – although algorithms which are specialized to (5) outperform such standard software; see Sect. 4.

The hope is, of course, that the solution of (5) coincides with the solution of (4) and with the original sparse vector x. Figure 3 provides an intuitive explanation why ℓ_1-minimization promotes sparse solutions. Here, $N = 2$ and $m = 1$, so one

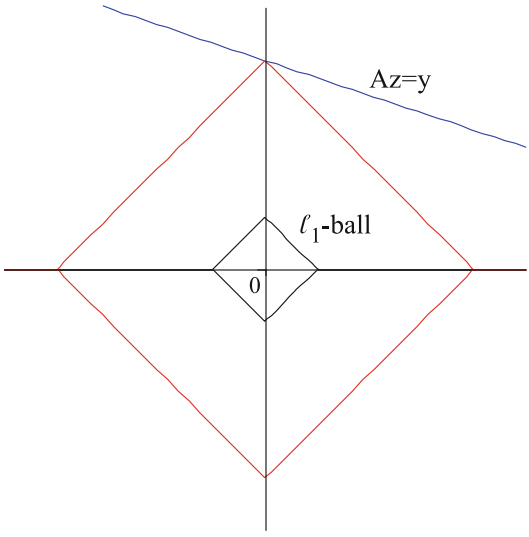


Fig. 3 The ℓ_1-minimizer within the affine space of solutions of the linear system $Az = y$ coincides with a sparsest solution

deals with a line of solutions $\mathcal{F}(y) = \{z : Az = y\}$ in $\mathbb{R}^2$. Except for pathological situations where $\ker A$ is parallel to one of the faces of the polytope B_1^2, there is a unique solution of the ℓ_1-minimization problem, which has minimal sparsity, i.e., only one nonzero entry.

Recovery results in the next sections make rigorous the intuition that ℓ_1-minimization indeed promotes sparsity.

For sparse recovery via greedy algorithms, one refers to the literature [151, 153].

The Null Space Property

The null space property is fundamental in the analysis of ℓ_1-minimization.

Definition 1. A matrix $A \in \mathbb{C}^{m\times N}$ is said to satisfy the *null space property* (NSP) of order k with constant $\gamma \in (0, 1)$ if

$$\|\eta_T\|_1 \leq \gamma \|\eta_{T^c}\|_1,$$

for all sets $T \subset \{1, \ldots, N\}$, $\#T \leq k$ and for all $\eta \in \ker A$.

The following sparse recovery result is based on this notion.

Theorem 1. *Let $A \in \mathbb{C}^{m\times N}$ be a matrix that satisfies the NSP of order k with constant $\gamma \in (0, 1)$. Let $x \in \mathbb{C}^N$ and $y = Ax$ and let x^* be a solution of the ℓ_1-minimization problem (5). Then*

$$\|x - x^*\|_1 \leq \frac{2(1+\gamma)}{1-\gamma}\sigma_k(x)_1. \tag{6}$$

In particular, if x is k-sparse, then $x^* = x$.

Proof. Let $\eta = x^* - x$. Then $\eta \in \ker A$ and

$$\|x^*\|_1 \leq \|x\|_1$$

because x^* is a solution of the ℓ_1-minimization problem (5). Let T be the set of the k-largest entries of x in absolute value. One has

$$\|x_T^*\|_1 + \|x_{T^c}^*\|_1 \leq \|x_T\|_1 + \|x_{T^c}\|_1.$$

It follows immediately from the triangle inequality that

$$\|x_T\|_1 - \|\eta_T\|_1 + \|\eta_{T^c}\|_1 - \|x_{T^c}\|_1 \leq \|x_T\|_1 + \|x_{T^c}\|_1.$$

Hence,

$$\|\eta_{T^c}\|_1 \leq \|\eta_T\|_1 + 2\|x_{T^c}\|_1 \leq \gamma\|\eta_{T^c}\|_1 + 2\sigma_k(x)_1,$$

or, equivalently,

$$\|\eta_{T^c}\|_1 \leq \frac{2}{1-\gamma}\sigma_k(x)_1. \tag{7}$$

Finally,

$$\|x - x^*\|_1 = \|\eta_T\|_1 + \|\eta_{T^c}\|_1 \leq (\gamma+1)\|\eta_{T^c}\|_1 \leq \frac{2(1+\gamma)}{1-\gamma}\sigma_k(x)_1$$

and the proof is completed. ■

One can also show that if all k-sparse x can be recovered from $y = Ax$ using ℓ_1-minimization, then necessarily A satisfies the NSP of order k with some constant $\gamma \in (0,1)$ [38, 87]. Therefore, the NSP is actually equivalent to sparse ℓ_1-recovery.

The Restricted Isometry Property

The NSP is somewhat difficult to show directly. The *restricted isometry property* (RIP) is easier to handle, and it also implies stability under noise as stated below.

Definition 2. The restricted isometry constant δ_k of a matrix $A \in \mathbb{C}^{m\times N}$ is the smallest number such that

$$(1-\delta_k)\|z\|_2^2 \leq \|Az\|_2^2 \leq (1+\delta_k)\|z\|_2^2, \tag{8}$$

for all $z \in \Sigma_k$.

A matrix A is said to satisfy the *restricted isometry property* of order k with constant δ_k if $\delta_k \in (0,1)$. It is easily seen that δ_k can be equivalently defined as

$$\delta_k = \max_{T\subset\{1,\dots,N\},\#T\leq k} \|A_T^* A_T - \mathrm{Id}\,\|_{2\to 2},$$

which means that *all* column submatrices of A with at most k columns are required to be well conditioned. The RIP implies the NSP as shown in the following lemma.

Lemma 2. *Assume that $A \in \mathbb{C}^{m\times N}$ satisfies the RIP of order $K = k + h$ with constant $\delta_K \in (0,1)$. Then A has the NSP of order k with constant $\gamma = \sqrt{\frac{k}{h}\frac{1+\delta_K}{1-\delta_K}}$.*

Proof. Let $\eta \in \mathcal{N} = \ker A$ and $T \subset \{1,\dots,N\}$, $\#T \leq k$. Define $T_0 = T$ and $T_1, T_2, \dots, T_s$ to be disjoint sets of indexes of size at most h, associated to a nonincreasing rearrangement of the entries of $\eta \in \mathcal{N}$, i.e.,

$$|\eta_j| \leq |\eta_i| \quad \text{for all } j \in T_\ell, i \in T_{\ell'}, \ell \geq \ell' \geq 1. \tag{9}$$

Note that $A\eta = 0$ implies $A\eta_{T_0\cup T_1} = -\sum_{j=2}^{s} A\eta_{T_j}$. Then, from the Cauchy–Schwarz inequality, the RIP, and the triangle inequality, the following sequence of inequalities is deduced:

$$\begin{aligned}\|\eta_T\|_1 &\le \sqrt{k}\|\eta_T\|_2 \le \sqrt{k}\|\eta_{T_0\cup T_1}\|_2 \\ &\le \sqrt{\frac{k}{1-\delta_K}}\|A\eta_{T_0\cup T_1}\|_2 = \sqrt{\frac{k}{1-\delta_K}}\|A\eta_{T_2\cup T_3\cup\cdots\cup T_s}\|_2 \\ &\le \sqrt{\frac{k}{1-\delta_K}}\sum_{j=2}^{s}\|A\eta_{T_j}\|_2 \le \sqrt{\frac{1+\delta_K}{1-\delta_K}}\sqrt{k}\sum_{j=2}^{s}\|\eta_{T_j}\|_2. \end{aligned} \tag{10}$$

It follows from (9) that $|\eta_i| \le |\eta_\ell|$ for all $i \in T_{j+1}$ and $\ell \in T_j$. Taking the sum over $\ell \in T_j$ first and then the ℓ_2-norm over $i \in T_{j+1}$ yields

$$|\eta_i| \le h^{-1}\|\eta_{T_j}\|_1, \quad \text{and} \quad \|\eta_{T_{j+1}}\|_2 \le h^{-1/2}\|\eta_{T_j}\|_1.$$

Using the latter estimates in (10) gives

$$\|\eta_T\|_1 \le \sqrt{\frac{1+\delta_K}{1-\delta_K}\frac{k}{h}}\sum_{j=1}^{s-1}\|\eta_{T_j}\|_1 \le \sqrt{\frac{1+\delta_K}{1-\delta_K}\frac{k}{h}}\|\eta_{T^c}\|_1, \tag{11}$$

and the proof is finished. ■

Taking $h = 2k$ above shows that $\delta_{3k} < 1/3$ implies $\gamma < 1$. By Theorem 1, recovery of all k-sparse vectors by ℓ_1-minimization is then guaranteed. Additionally, stability in ℓ_1 is also ensured. The next theorem shows that RIP implies also a bound on the reconstruction error in ℓ_2.

Theorem 2. *Assume $A \in \mathbb{C}^{m\times N}$ satisfies the RIP of order $3k$ with $\delta_{3k} < 1/3$. For $x \in \mathbb{C}^N$, let $y = Ax$ and x^* be the solution of the ℓ_1-minimization problem (5). Then*

$$\|x - x^*\|_2 \le C\frac{\sigma_k(x)_1}{\sqrt{k}}$$

with $C = \frac{2}{1-\gamma}\left(\frac{\gamma+1}{\sqrt{2}} + \gamma\right)$, $\gamma = \sqrt{\frac{1+\delta_{3k}}{2(1-\delta_{3k})}}$.

Proof. Similarly as in the proof of Lemma 2, denote $\eta = x^* - x \in \mathcal{N} = \ker A$, $T_0 = T$ the set of the $2k$-largest entries of η in absolute value, and T_j s of size at most k corresponding to the nonincreasing rearrangement of η. Then, using (10) and (11) with $h = 2k$ of the previous proof,

$$\|\eta_T\|_2 \leq \sqrt{\frac{1+\delta_{3k}}{2(1-\delta_{3k})}} k^{-1/2} \|\eta_{T^c}\|_1.$$

From the assumption $\delta_{3k} < 1/3$, it follows that $\gamma := \sqrt{\frac{1+\delta_{3k}}{2(1-\delta_{3k})}} < 1$. Lemmas 1 and 2 yield

$$\begin{aligned}\|\eta_{T^c}\|_2 = \sigma_{2k}(\eta)_2 &\leq (2k)^{-\frac{1}{2}} \|\eta\|_1 = (2k)^{-1/2} \left(\|\eta_T\|_1 + \|\eta_{T^c}\|_1\right) \\ &\leq (2k)^{-1/2} \left(\gamma \|\eta_{T^c}\|_1 + \|\eta_{T^c}\|_1\right) \leq \frac{\gamma+1}{\sqrt{2}} k^{-1/2} \|\eta_{T^c}\|_1.\end{aligned}$$

Since T is the set of $2k$-largest entries of η in absolute value, it holds

$$\|\eta_{T^c}\|_1 \leq \|\eta_{(\operatorname{supp} x_{[2k]})^c}\|_1 \leq \|\eta_{(\operatorname{supp} x_{[k]})^c}\|_1, \tag{12}$$

where $x_{[k]}$ is the best k-term approximation to x. The use of this latter estimate, combined with inequality (7), finally gives

$$\begin{aligned}\|x - x^*\|_2 \leq \|\eta_T\|_2 + \|\eta_{T^c}\|_2 &\leq \left(\frac{\gamma+1}{\sqrt{2}} + \gamma\right) k^{-1/2} \|\eta_{T^c}\|_1 \\ &\leq \frac{2}{1-\gamma} \left(\frac{\gamma+1}{\sqrt{2}} + \gamma\right) k^{-1/2} \sigma_k(x)_1.\end{aligned}$$

This concludes the proof. ■

The restricted isometry property implies also robustness under noise on the measurements. This fact was first noted in [26, 30].

Theorem 3. *Assume that the restricted isometry constant δ_{2k} of the matrix $A \in \mathbb{C}^{m \times N}$ satisfies*

$$\delta_{2k} < 1/\sqrt{2} \approx 0.7071 \tag{13}$$

Then the following holds for all $x \in \mathbb{C}^N$. Let noisy measurements $y = Ax + e$ be given with $\|e\|_2 \leq \eta$. Let x^* be the solution of

$$\min \|z\|_1 \quad \text{subject to } \|Az - y\|_2 \leq \eta. \tag{14}$$

Then

$$\|x - x^*\|_2 \leq C_1 \eta + C_2 \frac{\sigma_k(x)_1}{\sqrt{k}}$$

for some constants $C_1, C_2 > 0$ that depend only on δ_{2k}.

The constant in (13) was improved several times [22, 38, 74–76] until the present statement was reached in [19], which is actually optimal [45].

Coherence

The *coherence* is a by now classical way of analyzing the recovery abilities of a measurement matrix [50, 151]. For a matrix $A = (a_1|a_2|\cdots|a_N) \in \mathbb{C}^{m\times N}$ with normalized columns, $\|a_\ell\|_2 = 1$, it is defined as

$$\mu := \max_{\ell\neq k} |\langle a_\ell, a_k\rangle|.$$

Applying Gershgorin's disc theorem [93] to $A_T^* A_T - I$ with $\#T = k$ shows that

$$\delta_k \leq (k-1)\mu. \tag{15}$$

Several explicit examples of matrices are known which have small coherence $\mu = \mathcal{O}(1/\sqrt{m})$. A simple one is the concatenation $A = (I|F) \in \mathbb{C}^{m\times 2m}$ of the identity matrix and the unitary Fourier matrix $F \in \mathbb{C}^{m\times m}$ with entries $F_{j,k} = m^{-1/2} e^{2\pi i jk/m}$. It is easily seen that $\mu = 1/\sqrt{m}$ in this case. Furthermore, [143] gives several matrices $A \in \mathbb{C}^{m\times m^2}$ with coherence $\mu = 1/\sqrt{m}$. In all these cases, $\delta_k \leq C\frac{k}{\sqrt{m}}$. Combining this estimate with the recovery results for ℓ_1-minimization above shows that all k-sparse vectors x can be (stably) recovered from $y = Ax$ via ℓ_1-minimization provided

$$m \geq C' k^2. \tag{16}$$

At first sight, one might be satisfied with this condition since if k is very small compared to N, then still m might be chosen smaller than N and all k-sparse vectors can be recovered from the undersampled measurements $y = Ax$. Although this is great news for a start, one might nevertheless hope that (16) can be improved. In particular, one may expect that actually a linear scaling of m in k should be enough to guarantee sparse recovery by ℓ_1-minimization. The existence of matrices, which indeed provide recovery conditions of the form $m \geq Ck\log^\alpha(N)$ (or similar) with some $\alpha \geq 1$, is shown in the next section. Unfortunately, such results cannot be shown using simply the coherence because of the generally lower bound [143]

$$\mu \geq \sqrt{\frac{N-m}{m(N-1)}} \sim \frac{1}{\sqrt{m}} \qquad (N \text{ sufficiently large}).$$

In particular, it is not possible to overcome the "quadratic bottleneck" in (16) by using Gershgorin's theorem or Riesz–Thorin interpolation between $\|\cdot\|_{1\to 1}$ and $\|\cdot\|_{\infty\to\infty}$; see also [131, 141]. In order to improve on (16), one has to take into account also cancellations in the Gramian $A_T^* A_T - I$, and this task seems to be quite difficult using deterministic methods. Therefore, it will not come as a surprise that the major breakthrough in compressive sensing was obtained with random matrices.

It is indeed easier to deal with cancellations in the Gramian using probabilistic techniques.

RIP for Gaussian and Bernoulli Random Matrices

Optimal estimates for the RIP constants in terms of the number m of measurement matrices can be obtained for Gaussian, Bernoulli, or more general subgaussian random matrices.

Let X be a random variable. Then one defines a random matrix $A = A(\omega)$, $\omega \in \Omega$, as the matrix whose entries are independent realizations of X, where $(\Omega, \Sigma, \mathbb{P})$ is their common probability space. One assumes further that for any $x \in \mathbb{R}^N$ one has the identity $\mathbb{E}\|Ax\|_2^2 = \|x\|_2^2$, $\mathbb{E}$ denoting expectation.

The starting point for the simple approach in [7] is a concentration inequality of the form

$$\mathbb{P}\left(\left|\|Ax\|_2^2 - \|x\|_2^2\right| \geq \delta\|x\|_2^2\right) \leq 2e^{-c_0\delta^2 m}, \quad 0 < \delta < 1, \tag{17}$$

where $c_0 > 0$ is some constant.

The two most relevant examples of random matrices which satisfy the above concentration are the following:

1. **Gaussian matrices.** Here the entries of A are chosen as i.i.d. Gaussian random variables with expectation 0 and variance $1/m$. As shown in [1], Gaussian matrices satisfy (17).
2. **Bernoulli matrices** The entries of a Bernoulli matrices are independent realizations of $\pm 1/\sqrt{m}$ Bernoulli random variables, that is, each entry takes the value $+1/\sqrt{m}$ or $-1/\sqrt{m}$ with equal probability. Bernoulli matrices also satisfy the concentration inequality (17) [1].

Based on the concentration inequality (17), the following estimate on RIP constants can be shown [7, 26, 76, 110].

Theorem 4. *Assume $A \in \mathbb{R}^{m\times N}$ to be a random matrix satisfying the concentration property (17). Then there exists a constant C depending only on c_0 such that the restricted isometry constant of A satisfies $\delta_k \leq \delta$ with probability exceeding $1 - \varepsilon$ provided*

$$m \geq C\delta^{-2}(k\log(N/m) + \log(\varepsilon^{-1})).$$

Combining this RIP estimate with the recovery results for ℓ_1-minimization shows that all k-sparse vectors $x \in \mathbb{C}^N$ can be stably recovered from a random draw of A satisfying (17) with high probability provided

$$m \geq Ck\log(N/m). \tag{18}$$

Up to the logarithmic factor, this provides the desired linear scaling of the number m of measurements with respect to the sparsity k. Furthermore, as shown in Sect. 3 below, condition (18) cannot be further improved; in particular, the log-factor cannot be removed.

It is useful to observe that the concentration inequality is invariant under unitary transforms. Indeed, suppose that z is not sparse with respect to the canonical basis but with respect to a different orthonormal basis. Then $z = Ux$ for a sparse x and a unitary matrix $U \in \mathbb{C}^{N\times N}$. Applying the measurement matrix A yields

$$Az = AUx,$$

so that this situation is equivalent to working with the new measurement matrix $A' = AU$ and again sparsity with respect to the canonical basis. The crucial point is that A' satisfies again the concentration inequality (17) once A does. Indeed, choosing $x = U^{-1}x'$ and using unitarity gives

$$\begin{aligned}\mathbb{P}\left(\left|\|AUx\|_2^2 - \|x\|_2^2\right| \geq \delta\|x\|_{\ell_2^N}^2\right) &= \mathbb{P}\left(\left|\|Ax'\|_2^2 - \|U^{-1}x'\|_2^2\right| \geq \delta\|U^{-1}x'\|_{\ell_2^N}^2\right)\\ &= \mathbb{P}\left(\left|\|Ax'\|_2^2 - \|x'\|_2^2\right| \geq \delta\|x'\|_{\ell_2^N}^2\right) \leq 2e^{-c_0\delta^{-2}m}.\end{aligned}$$

Hence, Theorem 4 also applies to $A' = AU$. This fact is sometimes referred to as the *universality* of the Gaussian or Bernoulli random matrices. It does not matter in which basis the signal x is actually sparse. At the coding stage, where one takes random measurements $y = Az$, knowledge of this basis is not even required. Only the decoding procedure needs to know U.

Random Partial Fourier Matrices

While Gaussian and Bernoulli matrices provide optimal conditions for the minimal number of required samples for sparse recovery, they are of somewhat limited use for practical applications for several reasons. Often the application imposes physical or other constraints on the measurement matrix, so that assuming A to be Gaussian may not be justifiable in practice. One usually has only limited freedom to inject randomness in the measurements. Furthermore, Gaussian or Bernoulli matrices are not structured, so there is no fast matrix-vector multiplication available which may speed up recovery algorithms, such as the ones described in Sect. 4. Thus, Gaussian random matrices are not applicable in large-scale problems.

A very important class of structured random matrices that overcomes these drawbacks are random partial Fourier matrices, which were also the object of study in the very first papers on compressive sensing [26, 29, 128, 129]. A random partial Fourier matrix $A \in \mathbb{C}^{m\times N}$ is derived from the discrete Fourier matrix $F \in \mathbb{C}^{N\times N}$ with entries

$$F_{j,k} = \frac{1}{\sqrt{N}} e^{2\pi jk/N},$$

by selecting m rows uniformly at random among all N rows. Taking measurements of a sparse $x \in \mathbb{C}^N$ corresponds then to observing m of the entries of its discrete Fourier transform $\hat{x} = Fx$. It is important to note that the fast Fourier transform may be used to compute matrix-vector multiplications with A and A^* with complexity $\mathcal{O}(N \log(N))$. The following theorem concerning the RIP constant was proven in [131] and improves slightly on the results in [26, 129, 138].

Theorem 5. *Let $A \in \mathbb{C}^{m \times N}$ be the random partial Fourier matrix as just described. Then the restricted isometry constant of the rescaled matrix $\sqrt{\frac{N}{m}} A$ satisfies $\delta_k \leq \delta$ with probability at least $1 - N^{-\gamma \log^3(N)}$ provided*

$$m \geq C \delta^{-2} k \log^4(N). \tag{19}$$

The constants $C, \gamma > 1$ are universal.

Combining this estimate with the ℓ_1-minimization results above shows that recovery with high probability can be ensured for all k-sparse x provided

$$m \geq C k \log^4(N).$$

The plots in Fig. 1 illustrate an example of successful recovery from partial Fourier measurements.

The proof of the above theorem is not straightforward and involves Dudley's inequality as a main tool [131, 138]. Compared to the recovery condition (18) for Gaussian matrices, one suffers a higher exponent at the log-factor, but the linear scaling of m in k is preserved. Also a nonuniform recovery result for ℓ_1-minimization is available [29, 128, 131], which states that each k-sparse x can be recovered using a random draw of the random partial Fourier matrix A with probability at least $1-\varepsilon$ provided $m \geq C k \log(N/\varepsilon)$. The difference to the statement in Theorem 5 is that for each sparse x, recovery is ensured with high probability for a new random draw of A. It does not imply the existence of a matrix which allows recovery of *all* k-sparse x simultaneously. The proof of such recovery results does not make use of the restricted isometry property or the null space property.

One may generalize the above results to a much broader class of structured random matrices which arise from random sampling in bounded orthonormal systems. The interested reader is referred to [128, 129, 131, 132].

Another class of structured random matrices, for which recovery results are known, consist of partial random circulant and Toeplitz matrices. These correspond to subsampling the convolution of x with a random vector b at m fixed (deterministic) entries. The reader is referred to [130, 131, 133] for detailed information. Near-optimal estimates of the RIP constants of such type of random matrices have been established in [101]. Further types of random measurement matrices are discussed in [101, 122, 124, 137, 154]; see also [99] for an overview.

Compressive Sensing and Gelfand Widths

In this section a quite general viewpoint is taken. The question is investigated how well any measurement matrix and any reconstruction method – in this context usually called the *decoder* – may perform. This leads to the study of *Gelfand widths*, already mentioned in Sect. 2. The corresponding analysis allows to draw the conclusion that Gaussian random matrices in connection with ℓ_1-minimization provide optimal performance guarantees.

Following the tradition of the literature in this context, only the real-valued case will be treated. The complex-valued case is easily deduced from the real-valued case by identifying $\mathbb{C}^N$ with $\mathbb{R}^{2N}$ and by corresponding norm equivalences of ℓ_p-norms.

The measurement matrix $A \in \mathbb{R}^{m\times N}$ is here also referred to as the *encoder*. The set $\mathcal{A}_{m,N}$ denotes all possible encoder/decoder pairs (A, Δ) where $A \in \mathbb{R}^{m\times N}$ and $\Delta : \mathbb{R}^m \to \mathbb{R}^N$ is any (nonlinear) function. Then, for $1 \le k \le N$, the reconstruction errors over subsets $K \subset \mathbb{R}^N$, where $\mathbb{R}^N$ is endowed with a norm $\|\cdot\|_X$, are defined as

$$\sigma_k(K)_X := \sup_{x\in K} \sigma_k(x)_X,$$

$$E_m(K, X) := \inf_{(A,\Delta)\in A_{m,N}} \sup_{x\in K} \|x - \Delta(Ax)\|_X.$$

In words, $E_n(K, X)$ is the worst reconstruction error for the best pair of encoder/decoder. The goal is to find the largest k such that

$$E_m(K, X) \le C_0\sigma_k(K)_X.$$

Of particular interest for compressive sensing are the unit balls $K = B_p^N$ for $0 < p \le 1$ and $X = \ell_2^N$ because the elements of B_p^N are well approximated by sparse vectors due to Lemma 1. The proper estimate of $E_m(K, X)$ turns out to be linked to the geometrical concept of *Gelfand width*.

Definition 3. Let K be a compact set in a normed space X. Then the *Gelfand width* of K of order m is

$$d^m(K, X) := \inf_{\substack{Y \subset X \\ \mathrm{codim}(Y) \le m}} \sup\{\|x\|_X : x \in K \cap Y\},$$

where the infimum is over all linear subspaces Y of X of codimension less or equal to m.

The following fundamental relationship between $E_m(K, X)$ and the Gelfand widths holds.

Proposition 1. *Let $K \subset \mathbb{R}^N$ be a closed compact set such that $K = -K$ and $K + K \subset C_0 K$ for some constant C_0. Let $X = (\mathbb{R}^N, \|\cdot\|_X)$ be a normed space. Then*

$$d^m(K, X) \le E_m(K, X) \le C_0 d^m(K, X).$$

Proof. For a matrix $A \in \mathbb{R}^{m\times N}$, the subspace $Y = \ker A$ has codimension less or equal to m. Conversely, to any subspace $Y \subset \mathbb{R}^N$ of codimension less or equal to m, a matrix $A \in \mathbb{R}^{m\times N}$ can be associated, the rows of which form a basis for $Y^\perp$. This identification yields

$$d^m(K, X) = \inf_{A\in\mathbb{R}^{m\times N}} \sup\{\|\eta\|_X : \eta \in \ker A \cap K\}.$$

Let (A, Δ) be an encoder/decoder pair in $\mathcal{A}_{m,N}$ and $z = \Delta(0)$. Denote $Y = \ker(A)$. Then with $\eta \in Y$ also $-\eta \in Y$, and either $\|\eta - z\|_X \ge \|\eta\|_X$ or $\|-\eta - z\|_X \ge \|\eta\|_X$. Indeed, if both inequalities were false, then

$$\|2\eta\|_X = \|\eta - z + z + \eta\|_X \le \|\eta - z\|_X + \|-\eta - z\|_X < 2\|\eta\|_X,$$

a contradiction. Since $K = -K$, it follows that

$$\begin{aligned} d^m(K, X) &= \inf_{A\in\mathbb{R}^{m\times N}} \sup\{\|\eta\|_X : \eta \in Y \cap K\} \le \sup_{\eta\in Y\cap K} \|\eta - z\|_X \\ &= \sup_{\eta\in Y\cap K} \|\eta - \Delta(A\eta)\|_X \le \sup_{x\in K} \|x - \Delta(Ax)\|_X. \end{aligned}$$

Taking the infimum over all $(A, \Delta) \in \mathcal{A}_{m,N}$ yields

$$d^m(K, X) \le E_m(K, X).$$

To prove the converse inequality, choose an optimal Y such that

$$d^m(K, X) = \sup\{\|x\|_X : x \in Y \cap K\}.$$

(An optimal subspace Y always exists [107].) Let A be a matrix whose rows form a basis for $Y^\perp$. Denote the affine solution space $\mathcal{F}(y) := \{x : Ax = y\}$. One defines then a decoder as follows. If $\mathcal{F}(y) \cap K \ne \emptyset$, then choose some $\bar{x}(y) \in \mathcal{F}(y)$ and set $\Delta(y) = \bar{x}(y)$. If $\mathcal{F}(y) \cap K = \emptyset$, then $\Delta(y) \in \mathcal{F}(y)$. The following chain of inequalities is then deduced:

$$\begin{aligned} E_m(K, X) &\le \sup_y \sup_{x,x'\in\mathcal{F}(y)\cap K} \|x - x'\|_X \\ &\le \sup_{\eta\in C_0(Y\cap K)} \|\eta\|_X \le C_0 d^m(K, X), \end{aligned}$$

which concludes the proof. ∎

The assumption $K + K \subset C_0 K$ clearly holds for norm balls with $C_0 = 2$ and for quasi-norm balls with some $C_0 \geq 2$. The next theorem provides a two-sided estimate of the Gelfand widths $d^m(B_p^N, \ell_2^N)$ [48, 77, 157]. Note that the case $p = 1$ was considered much earlier in [77, 79, 96].

Theorem 6. *Let $0 < p \leq 1$. There exist universal constants $C_p, D_p > 0$ such that the Gelfand widths $d^m(B_p^N, \ell_2^N)$ satisfy*

$$C_p \min\left\{1, \frac{\ln(2N/m)}{m}\right\}^{1/p-1/2} \leq d^m(B_p^N, \ell_2^N) \leq D_p \min\left\{1, \frac{\ln(2N/m)}{m}\right\}^{1/p-1/2} \tag{20}$$

Combining Proposition 1 and Theorem 6 gives in particular, for large m,

$$\tilde{C}_1 \sqrt{\frac{\log(2N/m)}{m}} \leq E_m(B_1^N, \ell_2^N) \leq \tilde{D}_1 \sqrt{\frac{\log(2N/m)}{m}}. \tag{21}$$

This estimate implies a lower estimate for the minimal number of required samples which allows for approximate sparse recovery using any measurement matrix and any recovery method whatsoever. The reader should compare the next statement with Theorem 2.

Corollary 1. *Suppose that $A \in \mathbb{R}^{m \times N}$ and $\Delta : \mathbb{R}^m \to \mathbb{R}^N$ such that*

$$\|x - \Delta(Ax)\|_2 \leq C \frac{\sigma_k(x)_1}{\sqrt{k}}$$

for all $x \in B_1^N$ and some constant $C > 0$. Then necessarily

$$m \geq C' k \log(2N/m). \tag{22}$$

Proof. Since $\sigma_k(x)_1 \leq \|x\|_1 \leq 1$, the assumption implies $E_m(B_1^N, \ell_2^N) \leq C k^{-1/2}$. The lower bound in (21) combined with Proposition 1 yields

$$\tilde{C}_1 \sqrt{\frac{\log(2N/m)}{m}} \leq E_m(B_1^N, \ell_2^N) \leq C k^{-1/2}.$$

Consequently, $m \geq C' k \log(eN/m)$ as claimed. ∎

In particular, the above lemma applies to ℓ_1-minimization and consequently $\delta_k \leq 0.5$ (say) for a matrix $A \in \mathbb{R}^{m \times N}$ implies $m \geq C k \log(N/m)$. Therefore, the recovery results for Gaussian or Bernoulli random matrices with ℓ_1-minimization stated above are optimal.

It can also be shown that a stability estimate in the ℓ_1-norm of the form $\|x - \Delta(Ax)\|_1 \leq C\sigma_k(x)_1$ for all $x \in \mathbb{R}^N$ implies (22) as well [46, 77].

Extensions of Compressive Sensing

The problem of recovering nearly sparse vectors from a minimal number of nonadaptive linear measurements can be extended in several forms. In the following two possible, mutually related, extensions are addressed.

The first refers to the object to be recovered: while in the theory of compressive sensing presented so far nearly sparse vectors are the unknowns of the problem at hand, one may consider as well matrices of approximate low rank. This leads to the problem of finding a matrix of minimal rank consistent with a given underdetermined linear system of equations.

The second generalization of compressive sensing considers nonlinear nonadaptive measurements. The simplest nonlinearity consists in *quadratic* measurements, given by the squared magnitude of the scalar products of the vector with respect to a fixed basis or a frame. The associated recovery task – called the *phase retrieval problem* – appears in many physical situations where only intensity values can be observed. More generally, one may consider the recovery from higher-order measurements. Surprisingly enough, polynomial-type, and in particular quadratic, measurements of vectors can be recast into an affine low-rank minimization problem discussed above, creating a connection between these two extensions of compressive sensing.

Affine Low-Rank Minimization

Here one denotes the space of real or complex $n \times p$ matrices by $M_{n\times p}$. Given a linear map $\mathscr{S} : M_{n\times p} \to \mathbb{C}^m$, with $m \ll pn$, and a vector $y \in \mathbb{C}^m$, one considers the affine rank minimization problem

$$\min_{X \in M_{n\times p}} \operatorname{rank}(X) \quad \text{subject to } \mathscr{S}(X) = y. \tag{23}$$

An important special case of low-rank matrix recovery is matrix completion, where $\mathscr{S}$ samples a few entries,

$$\mathscr{S}(X)_\ell = x_{ij}, \tag{24}$$

for some i, j depending on ℓ.

Like for the ℓ_0-minimization problem (4), the affine rank minimization problem is NP-hard in general [76, 114, 135]; therefore, it is desirable to have tractable alternatives. In [67], Fazel studied *nuclear norm* minimization for this purpose. Here, the nuclear norm $\|X\|_*$ of a matrix X is the ℓ_1-norm of its singular values and is the largest convex envelope of the rank function [67]. One solves then

$$\min_{X \in M_{n\times p}} \|X\|_* \quad \text{subject to } \mathscr{S}(X) = y. \tag{25}$$

There are two known regimes where nuclear norm minimization can be guaranteed to return minimal-rank solutions:

RIP measurement maps

The (rank) restricted isometry property is analogous to the classical restricted isometry property from compressive sensing already mentioned in Sect. 3:

Definition 4 (Rank-Restricted Isometry Property [135]). Let $\mathscr{S} : M_{n\times p} \to \mathbb{C}^m$ be a linear map. For an integer k, with $1 \le k \le n$, define the k-restricted isometry constant $\delta_k = \delta_k(\mathscr{S}) > 0$ to be the smallest number such that

$$(1-\delta_k)\|X\|_F^2 \le \|\mathscr{S}(X)\|_{\ell_2^m}^2 \le (1+\delta_k)\|X\|_F^2$$

holds for all k-rank matrices X.

It is known that nuclear norm minimization (25) recovers all matrices X of rank at most k from the measurements $y = \mathscr{S}(X)$, provided $\delta_{2k} < 1/\sqrt{2}$; see [19, 24, 76, 121, 135].

The restricted isometry property is known to hold for Gaussian (or more generally subgaussian) measurement maps [24, 135], which are of the form

$$\mathscr{S}(X)_\ell = \sum_{k,j} a_{\ell,k,j} X_{k,j}, \quad \ell = 1, \dots, m, \tag{26}$$

where the $a_{\ell,k,j}$ are independent normal distributed random variables with mean zero and variance $1/m$. Such a map satisfies $\delta_k \le \delta \in (0,1)$ with high probability provided

$$m \ge C_\delta \max\{p,n\}k; \tag{27}$$

see Theorem 2.3 in [24]. Since the degrees of freedom of a rank k matrix $X \in M_{n\times p}$ are $k(n+p-k)$, the above bound matches this number up to possibly a constant. Therefore, the bound (27) is optimal.

It follows from recent results in [3, 100] that the restricted isometry property also holds for certain *structured* random maps if slightly more measurements are allowed. In particular, let $\mathscr{S} = P_\Omega FD$, where $D : M_{n\times p} \to M_{n\times p}$ performs independent random sign flips of all entries of a matrix, $F : M_{n\times p} \to M_{n\times p}$ represents the (suitably normalized) $2D$ Fourier transform, and $P_\Omega : M_{n\times p} \to \mathbb{C}^m$ is the coordinate projection which extracts the entries of the matrix on a random set $\Omega \subset [n]\times[p]$ which is chosen uniformly at random among all subsets of cardinality m. Then the rank-restricted isometry constants of $\mathscr{S}$ satisfy $\delta_k \le \delta \in (0,1)$ with high probability as long as

$$m \ge C_\delta \max\{p,n\}k \log^5(pn). \tag{28}$$

This follows from recent findings in [3, 100] for which such random partial Fourier measurements satisfy a concentration inequality of the form

$$\mathbb{P}\Big(\big| \|\mathscr{S}(X)\|^2 - \|X\|_F^2 \big| \geq \varepsilon \|X\|_F^2 \Big) \leq 2 \exp\left(-\frac{m}{2} C_\varepsilon \log^{-4}(pn)\right), \tag{29}$$

for $\varepsilon \in (0, 1)$, together with the same proof strategy employed for subgaussian measurement maps based on covering arguments [24, 76]. This was first noted in the introduction of [73].

Matrix Completion

In the matrix completion setup (24) where measurements are pointwise observations of entries of the matrix, there are obvious rank one matrices in the kernel of the operator $\mathscr{S}$; therefore, the RIP fails completely in this setting, and "localized" low-rank matrices in the null space of $\mathscr{S}$ cannot be recovered by any method whatsoever. However, if certain conditions on the left and right singular vectors of the underlying low-rank matrix are imposed, essentially requiring that such vectors are uncorrelated with the canonical basis, then it was shown in [25, 27, 134] that such incoherent matrices of rank at most k can be recovered from m randomly chosen entries with high probability provided

$$m \geq C k \max\{n, p\} \log^2(\max\{n, p\}).$$

Up to perhaps the exponent 2 at the log-term, this is optimal. One refers to [25, 27, 134] for detailed statements. In [88, 89] extensions to quantum state tomography can be found.

Nonlinear Measurements

Phase Retrieval

In order to understand how one can generalize compressive sensing towards nonlinear measurements, it is instructive to start with the *phase retrieval* problem. Here, the measurements of a signal x are given by $y_i = |\langle a_i, x\rangle|^2$ for some vectors a_i, $i = 1, \ldots, m$ and the task is to reconstruct x up to a global phase factor. Practically successful approaches can be found in the physics and optimization literature for well over a decade; see, for instance, [8, 68]. In these investigations, however, neither bounds of the minimal number nor a theoretical discussion on the type of measurements is provided for the phase retrieval problem to ensure reconstruction. Moreover, there are no guarantees for these algorithms to converge to the expected solution. Only recently, uniqueness of feasible solutions was provided by Balan et al. [5] using methods from algebraic geometry. The authors proved injectivity for signals $x \in \mathbb{R}^N$ and $m \geq 2N - 1$ generic measurements or for complex signals $x \in \mathbb{C}^N$ and $m \geq 4N - 2$ generic measurements. Unfortunately, this theoretical work did not provide recovery guarantees for a practical algorithm. The main tool in this work was the observation that quadratic measurements can be written as

$$y_i = |\langle a_i, x\rangle|^2 = \langle a_i a_i^*, xx^*\rangle, \quad i = 1, \dots, m, \tag{30}$$

where the inner product on the right-hand side here denotes the Hilbert-Schmidt inner product $\langle A, B\rangle = \mathrm{Tr}(AB^*)$. Consequently x can be identified as the unique solution of the phase retrieval problem (up to a global phase factor) if $X = xx^*$ is the unique rank one (and thus the minimum rank!) positive semidefinite solution of the linear equations

$$y_i = \mathscr{S}(X)_i = \langle a_i a_i^*, X\rangle, \quad i = 1, \dots, m. \tag{31}$$

This well-known lifting trick is also the basis of the PhaseLift algorithm by Candès et al. [32]. Here, the main ingredient is to pass from the rank minimization problem constrained by (31) to its convex relaxation (25). Using techniques from random matrix theory, the authors showed recovery guarantees for $m = O(N \log(N))$ Gaussian measurements under no structural (sparsity) assumptions on the signal. Although PhaseLift is often not considered a very efficient approach (see Sect. 4 for possible implementations) as one needs to solve a semidefinite program for $N \times N$ matrices – thus squaring the dimension – this work is considered groundbreaking, as for the first time it provided recovery guarantees for a polynomial time algorithm. The number of Gaussian measurements for PhaseLift to recover a generic signal has later been improved to $m = O(N)$ in [23] and for a variant of PhaseLift to $m = O(k^2 \log(N))$ for k-sparse signals in [105]. These recovery guarantees include also stability with respect to measurement noise.

The question of how many measurements are necessary to allow for stable phase retrieval, independently of the algorithm used, has been addressed in a very general setting by Eldar and Mendelson [64]. They showed that the necessary number of phaseless measurements to allow for the recovery of a signal x which is known to lie in a given set S can be estimated in terms of the so-called Gaussian width of the set S. This result is not restricted to Gaussian measurements but extends also to measurement vectors with independent subgaussian entries which in addition satisfy a small ball probability bound. Building upon this work, Ehler et al. [60] showed recovery guarantees for certain greedy algorithms for complex measurements of the same type. Other algorithms, for which recovery guarantees have been provided for Gaussian measurements, include polarization-based approaches [4] and alternating minimization [116].

Rigorous recovery guarantees for phase retrieval have been extended to randomly masked Fourier transforms in [31, 90] and to spherical designs in [91]. These works address more realistic measurement scenarios than Gaussian measurements for practical applications such as X-ray crystallography; see also [99].

Higher-Order Polynomial Measurements

Building upon the successful experience on phase retrieval problems, compressed sensing theory can be further extended to solving nonlinear problems of the type

$$\min \|z\|_0 \quad \text{subject to } y_i = A_i(z), \quad i = 1, \dots, m, \tag{32}$$

where $A_i : \mathbb{C}^N \to \mathbb{C}, i = 1, \ldots, m$, are smooth functions. For q being a sufficiently large even integer, one approximates A_i by its qth-order Taylor expansion,

$$A_i(x) \approx \sum_{0 \le |\alpha| \le q} \frac{(x - x_0)^\alpha}{\alpha!} \partial^\alpha A_i(x_0) = \bar{x}^* \mathcal{A}_i \bar{x},$$

where $\mathcal{A}_i$ is a $\binom{N + \frac{q}{2}}{\frac{q}{2}} \times \binom{N + \frac{q}{2}}{\frac{q}{2}}$-symmetric matrix and $\bar{x}$ is the vector whose entries are all the monomials of the elements of x with degree less or equal to $q/2$. Hereby, the standard multi-index notation is used.

Example 1. Let $x = (x_1, x_2)^T$ and $A(x) = 1 + x_1 + x_2^3$. The 4th-order Taylor expansion around $x_0 = 0$ is given by

$$A(x) = \bar{x}^* \mathcal{A} \bar{x}$$

where $\bar{x} = (1, x_1, x_2, x_1 x_2, x_1^2, x_2^2)^T$ and

$$\mathcal{A} = \begin{pmatrix} 1 & 1/2 & 0 & 0 & 0 & 0 \\ 1/2 & 0 & 0 & 0 & 1/12 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1/12 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

One observes again that

$$y_i = \bar{x}^* \mathcal{A}_i \bar{x} = \langle \mathcal{A}_i, \bar{x}\bar{x}^* \rangle, \quad i = 1, \ldots, m, \tag{33}$$

where the inner product on the right-hand side denotes the Hilbert-Schmidt inner product. Consequently x can be recognized as the solution of (32) if $X = \bar{x}\bar{x}^*$ is the unique rank one positive definite solution of the linear equations

$$y_i = \mathscr{S}(X)_i = \langle \mathcal{A}_i, X \rangle, \quad i = 1, \ldots, m. \tag{34}$$

As for the PhaseLift algorithm for phase retrieval problems, one considers again instead of the rank minimization problem constrained by (34) its convex relaxation (25). One also notices that in case the vector x is sparse, then X is sparse in addition to being of rank one. Hence, one can combine the nuclear norm with the classical ℓ_1-norm penalization and consider the problem

$$\min_{X \in M_{p \times p}, 0 \preceq X} \|X\|_* + \gamma \|X\|_{\ell_1}, \text{ subject to } \mathscr{S}(X) = y, \tag{35}$$

where $p = \binom{N + \frac{q}{2}}{\frac{q}{2}}$ and $\gamma > 0$ is a suitable parameter. One refers to (35) as *nonlinear basis pursuit*. This algorithm works surprisingly well in practice to recover sparse vectors from very few measurements [118], but theoretical guarantees for stable and unique recovery are open to date.

Applications

Compressive sensing can be potentially used in all applications where the task is the reconstruction of a signal or an image from linear measurements, while taking many of those measurements – in particular, a complete set of measurements – is a costly, lengthy, difficult, dangerous, impossible, or otherwise undesired procedure. Additionally, there should be reasons to believe that the signal is sparse in a suitable basis (or frame). Empirically, the latter applies to most types of signals.

In computerized tomography, for instance, one would like to obtain an image of the inside of a human body by taking X-ray images from different angles. Taking an almost complete set of images would expose the patient to a large and dangerous dose of radiation, so the amount of measurements should be as small as possible and nevertheless guarantee a good enough image quality. Such images are usually nearly piecewise constant and therefore nearly sparse in the gradient, so there is a good reason to believe that compressive sensing is well applicable. And indeed, it is precisely this application that started the investigations on compressive sensing in the seminal paper [29].

Also radar imaging seems to be a very promising application of compressive sensing techniques [65, 66, 94, 144]. One is usually monitoring only a small number of targets, so that sparsity is a very realistic assumption. Standard methods for radar imaging actually also use the sparsity assumption, but only at the very end of the signal processing procedure in order to clean up the noise in the resulting image. Using sparsity systematically from the very beginning by exploiting compressive sensing methods is therefore a natural approach.

Further potential applications include wireless communication [147], astronomical signal and image processing [14], analog to digital conversion [154], camera design [58], and imaging [136, 159, 160].

Affine low-rank minimization problems (23) arise in many areas of science and technology, including system identification [113], collaborative filtering, quantum state tomography [88, 89], signal, and image processing. An important special case is the matrix completion problem [25, 27, 134], where the task consists in filling in missing entries of a large low-rank data matrix. For applications of low-rank matrix recovery in quantum state tomography, one can refer to [88, 89]. Applications of phase retrieval include diffraction imaging and X-ray crystallography; see, for instance, [111].

4 Numerical Methods

The previous sections showed that ℓ_1-minimization performs very well in recovering sparse or approximately sparse vectors from undersampled measurements. In applications, it is important to have fast methods for actually solving ℓ_1-minimization problems. Two such methods – Chambolle and Pock's primal-dual algorithm [34] and iteratively re-weighted least squares (IRLS) [44] – will be explained in more detail below.

As a first remark, the ℓ_1-minimization problem

$$\min \|x\|_1 \quad \text{subject to } Ax = y \tag{36}$$

is in the real case equivalent to the linear program

$$\min \sum_{j=1}^{2N} v_j \quad \text{subject to} \quad v \geq 0,\ (A|-A)v = y. \tag{37}$$

The solution x^* to (36) is obtained from the solution v^* of (37) via $x^* = (\mathrm{Id}\,|\,-\mathrm{Id})v^*$. Any linear programming method may therefore be used for solving (36). The simplex method and interior point methods apply in particular [115], and standard software may be used. (In the complex case, (36) is equivalent to a second-order cone program (SOCP) and can also be solved with interior point methods.) However, such methods and software are of general purpose, and one may expect that methods specialized to (36) outperform such existing standard methods. Moreover, standard software often has the drawback that one has to provide the full matrix rather than fast routines for matrix-vector multiplication which are available, for instance, in the case of partial Fourier matrices. In order to obtain the full performance of such methods, one would therefore need to reimplement them, which is a daunting task because interior point methods usually require much fine-tuning. On the contrary, the two specialized methods described below are rather simple to implement and very efficient. Many more methods are available nowadays, including the homotopy and LARS method [55, 59, 119, 120]; greedy methods; such as orthogonal matching pursuit [151], CoSaMP [153]; and iterative hard thresholding [13, 69], which may offer better complexity than standard interior point methods. Due to space limitations, however, only the two methods below are explained in detail.

A Primal-Dual Algorithm

The reconstruction approaches discussed in the previous section can often be solved by optimization problems of the form

$$\min_{x\in\mathbb{C}^N} F(Ax) + G(x) \tag{38}$$

with A being an $m \times N$ matrix and F, G being convex functions with values in $(-\infty, \infty]$. For instance, the ℓ_1-minimization problem (5) can be recast as (38) with $G(x) = \|x\|_1$ and $F(z) = \chi_{\{y\}}(z)$, where $\chi_{\{y\}}$ is the characteristic function of the singleton set $\{y\}$ that takes the value 0 on y and ∞ elsewhere. Moreover, (14) is equivalent to (38) with $G(x) = \|x\|_1$ and $F = \chi_{\{w:\|y-w\|_2\le\eta\}}$. Also the nuclear norm minimization problem (25) takes the form (38) with G being the nuclear norm; see also below.

The so-called dual problem [17, 76] of (38) is given by

$$\max_{\xi\in\mathbb{C}^m} -F^*(\xi) - G^*(-A^*\xi), \tag{39}$$

where

$$F^*(\xi) = \sup\{\mathrm{Re}\langle z, \xi\rangle - F(z) : z \in \mathbb{C}^m\}$$

is the convex (Fenchel) conjugate of F and likewise G^* is the convex conjugate of G. The joint primal-dual optimization is equivalent to solving the saddle point problem

$$\min_{x\in\mathbb{C}^N} \max_{\xi\in\mathbb{C}^m} \mathrm{Re}\langle Ax, \xi\rangle + G(x) - F^*(\xi). \tag{40}$$

The primal-dual algorithm introduced by Chambolle and Pock in [34], generalizing [126], and described below solves this problem iteratively. In order to formulate it, one needs to introduce the so-called proximal mapping (proximity operator). For the convex function G, it is defined as

$$P_G(\tau, z) := \operatorname*{argmin}_{x\in\mathbb{C}^N} \left\{ \tau G(x) + \frac{1}{2}\|x - z\|_2^2 \right\}.$$

For $G(x) = \|x\|_1$, the proximal mapping reduces to the well-known soft-thresholding operator, which is defined componentwise as

$$S_\tau(z)_\ell = \begin{cases} \mathrm{sgn}(z_\ell)(|z_\ell| - \tau) & \text{if } |z_\ell| \ge \tau, \\ 0 & \text{otherwise.} \end{cases}$$

One also requires below the proximal mapping associated to the convex conjugate F^*.

The primal-dual algorithm performs an iteration on the dual variable, the primal variable, and an auxiliary primal variable. Starting from initial points $x^0, \bar{x}^0 = x \in \mathbb{C}^N, \xi^0 \in \mathbb{C}^m$ and parameters $\tau, \sigma > 0$, $\theta \in [0, 1]$, one iteratively computes

$$\xi^{n+1} := P_{F^*}(\sigma; \xi^n + \sigma A \bar{x}^n), \tag{41}$$

$$x^{n+1} := P_G(\tau; x^n - \tau A^* \xi^{n+1}), \tag{42}$$

$$\bar{x}^{n+1} := x^{n+1} + \theta(x^{n+1} - x^n). \tag{43}$$

It can be shown that (x, ξ) is a fixed point of these iterations if and only if (x, ξ) is a saddle point of (40) which is equivalent to x being a solution of (38) and ξ of (39); see [76, Proposition 15.6]. For this reason, the algorithm can be interpreted as a fixed point iteration.

For the case that $\theta = 1$, the convergence of the primal-dual algorithm has been established in [34].

Theorem 7. *Consider the primal-dual algorithm* (41)–(43) *with* $\theta = 1$ *and* $\tau, \sigma > 0$ *such that* $\tau\sigma \|A\|_{2\to 2}^2 < 1$. *If the problem* (40) *has a saddle point, then the sequence* (x^n, ξ^n) *converges to a saddle point of* (40) *and in particular* (x^n) *converges to a minimizer of* (38).

One observes that for the ℓ_1 and nuclear norm minimization problems of this chapter, the assumption of the theorem that (40) has a saddle point will always be satisfied. An estimate of the convergence rate can be shown as well. One refers to [34] for details. Preconditioning was considered in [125], where also parameter choices are discussed. Convergence results for values of θ different from 1 have been obtained in [92] for slight modifications of the algorithm.

One considers again some special cases. For ℓ_1-minimization (36), where $G(x) = \|x\|_1$ and $F = \chi_{\{y\}}$, the proximal mapping P_G reduces to soft-thresholding (41) while the proximal mapping associated to $F^*(\xi) = \mathrm{Re}\langle y, \xi\rangle$ is given by

$$P_{F^*}(\sigma, \xi) = \xi - \sigma y.$$

Therefore, all operations required in the algorithm (41)–(43) are simple and easy to implement. Also, note that if fast matrix-vector multiplication algorithms are available for A and A^*, these can be easily exploited.

In the case of the nuclear norm minimization problem (25), where $G(X) = \|X\|_*$ and $F = \chi_{\{y\}}$, the proximal mapping of G is given by singular-value soft thresholding: If $Z = U\Sigma V^*$ is the singular value decomposition of a matrix Z with unitary matrices U and V and a diagonal matrix $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_n)$ with $\sigma_1 \geq \sigma_2 \geq \ldots \geq 0$, then

$$P_{\|\cdot\|_*}(\tau, Z) = U\,\mathrm{diag}(S_\tau(\sigma_1), \cdots, S_\tau(\sigma_n))V^*,$$

i.e., the soft-thresholding operator is applied to the singular values. With this information together with the proximal mapping of F^* given above, the iterations (41)–(43) can be implemented for nuclear norm minimization.

Of course, the described algorithm applies to many more optimization problems and turns out to be very efficient when the proximal mappings are simple to evaluate. Notice that the iterations (41)–(43) are strongly influenced by backward–forward splitting methods [41] for the primal problem (38) with differentiable F, which are of the form

$$x^n = P_G(\tau; x^n - \tau \nabla F(x^n)). \tag{44}$$

In fact, (41) can be considered a backward–forward step for the dual variable, while (42) is a backward–forward step for the primal variable. A particular instance of (44) is the soft-thresholding algorithm for minimizing functionals of the form $\|Ax - y\|_2^2 + \lambda \|x\|_1$; see [41, 43] and [9, 10] for an accelerated version. Related optimization algorithms include Douglas–Rachford splittings [39, 40, 57] and the alternating direction method of multipliers [40, 83]. One refers to [142, Chapter 7], [40], [76, Chapter 15], and [156] for overviews on these and further approaches.

Iteratively Re-weighted Least Squares

This section is concerned with an iterative algorithm which, under the condition that A satisfies the NSP (see Definition 1), is guaranteed to reconstruct vectors with the same error estimate (6) as ℓ_1-minimization. The following discussion is restricted to the real-valued case. This algorithm has a guaranteed linear rate of convergence which can even be improved to a superlinear rate with a small modification. First, a brief introduction aims at shedding light on the basic principles of this algorithm and their interplay with sparse recovery and ℓ_1-minimization.

Denote $\mathcal{F}(y) = \{x : Ax = y\}$ and $\mathcal{N} = \ker A$. The starting point is the trivial observation that $|t| = \frac{t^2}{|t|}$ for $t \neq 0$. Hence, an ℓ_1-minimization can be recasted into a weighted ℓ_2-minimization, with the hope that

$$\arg\min_{x \in \mathcal{F}(y)} \sum_{j=1}^N |x_j| \approx \arg\min_{x \in \mathcal{F}(y)} \sum_{j=1}^N x_j^2 |x_j^*|^{-1},$$

as soon as x^* is the desired ℓ_1-norm minimizer. The advantage of the reformulation consists in the fact that minimizing the smooth quadratic function t^2 is an easier task than the minimization of the nonsmooth function $|t|$. However, the obvious drawbacks are that neither one disposes of x^* a priori (this is the vector one is interested to compute!) nor one can expect that $x_j^* \neq 0$ for all $j = 1, \ldots, N$, since one hopes for k-sparse solutions.

Suppose one has a good approximation w_j^n of $|(x_j^*)^2 + \epsilon_n^2|^{-1/2} \approx |x_j^*|^{-1}$, for some $\epsilon_n > 0$. One computes

$$x^{n+1} = \arg\min_{x \in \mathcal{F}(y)} \sum_{j=1}^N x_j^2 w_j^n \tag{45}$$

and then updates $\epsilon_{n+1} \leq \epsilon_n$ by some rule to be specified later. Further, one sets

$$w_j^{n+1} = |(x_j^{n+1})^2 + \epsilon_{n+1}^2|^{-1/2} \tag{46}$$

and iterates the process. The hope is that a proper choice of $\epsilon_n \to 0$ allows the iterative computation of an ℓ_1-minimizer. The next sections investigate convergence of this algorithm and properties of the limit.

Weighted ℓ_2-Minimization

Suppose that the weight w is *strictly positive* which means that $w_j > 0$ for all $j \in \{1, \ldots, N\}$. Then $\ell_2(w)$ is a Hilbert space with the inner product

$$\langle u, v \rangle_w := \sum_{j=1}^{N} w_j u_j v_j . \tag{47}$$

Define

$$x^w := \arg \min_{z \in \mathcal{F}(y)} \|z\|_{2,w}, \tag{48}$$

where $\|z\|_{2,w} = \langle z, z \rangle_w^{1/2}$. Because the $\|\cdot\|_{2,w}$-norm is strictly convex, the minimizer x^w is necessarily unique; it is characterized by the orthogonality conditions

$$\langle x^w, \eta \rangle_w = 0, \quad \text{for all } \eta \in \mathcal{N}. \tag{49}$$

An Iteratively Re-weighted Least Squares Algorithm (IRLS)

An IRLS algorithm appears for the first time in the Ph.D. thesis of Lawson in 1961 [102], in the form of an algorithm for solving uniform approximation problems. This iterative algorithm is now well known in classical approximation theory as Lawson's algorithm. In [37] it is proved that it obeys a linear convergence rate. In the 1970s, extensions of Lawson's algorithm for ℓ_p-minimization, and in particular ℓ_1-minimization, were introduced. In signal analysis, IRLS was proposed as a technique to build algorithms for sparse signal reconstruction in [86]. The interplay of the NSP, ℓ_1-minimization, and a re-weighted least square algorithm has been clarified only recently in the work [44].

The analysis of the algorithm (45) and (46) starts from the observation that

$$|t| = \min_{w>0} \frac{1}{2} \left(w t^2 + w^{-1} \right),$$

the minimum being attained for $w = \frac{1}{|t|}$. Inspired by this simple relationship, given a real number $\epsilon > 0$ and a weight vector $w \in \mathbb{R}^N$, with $w_j > 0$, $j = 1, \ldots, N$, one introduces the functional

$$\mathcal{J}(z, w, \epsilon) := \frac{1}{2} \sum_{j=1}^{N} \left(z_j^2 w_j + \epsilon^2 w_j + w_j^{-1} \right), \quad z \in \mathbb{R}^N. \tag{50}$$

The algorithm roughly described in (45) and (46) can be recast as an alternating method for choosing minimizers and weights based on the functional $\mathcal{J}$. To describe this more rigorously, recall that $r(z)$ denotes the nonincreasing rearrangement of a vector $z \in \mathbb{R}^N$.

Algorithm IRLS. Initialize by taking $w^0 := (1, \dots, 1)$. Set $\epsilon_0 := 1$. Then recursively define, for $n = 0, 1, \dots,$

$$x^{n+1} := \arg \min_{z \in \mathcal{F}(y)} \mathcal{J}(z, w^n, \epsilon_n) = \arg \min_{z \in \mathcal{F}(y)} \|z\|_{2,w^n} \tag{51}$$

and

$$\epsilon_{n+1} := \min \left\{ \epsilon_n, \frac{r_{K+1}(x^{n+1})}{N} \right\}, \tag{52}$$

where K is a fixed integer that will be specified later. Set

$$w^{n+1} := \arg \min_{w>0} \mathcal{J}(x^{n+1}, w, \epsilon_{n+1}). \tag{53}$$

The algorithm stops if $\epsilon_n = 0$; in this case, define $x^j := x^n$ for $j > n$. In general, the algorithm generates an infinite sequence $(x^n)_{n \in \mathbb{N}}$ of vectors.

Each step of the algorithm requires the solution of a weighted least squares problem. In matrix form

$$x^{n+1} = D_n^{-1} A^* (A D_n^{-1} A^*)^{-1} y, \tag{54}$$

where D_n is the $N \times N$ diagonal matrix the j th diagonal entry of which is w_j^n. Once x^{n+1} is found, the weight w^{n+1} is given by

$$w_j^{n+1} = [(x_j^{n+1})^2 + \epsilon_{n+1}^2]^{-1/2}, \quad j = 1, \dots, N. \tag{55}$$

Convergence Properties

Lemma 3. *Set $L := \mathcal{J}(x^1, w^0, \epsilon_0)$. Then*

$$\|x^n - x^{n+1}\|_2^2 \le 2L \left[\mathcal{J}(x^n, w^n, \epsilon_n) - \mathcal{J}(x^{n+1}, w^{n+1}, \epsilon_{n+1}) \right].$$

Hence $(J(x^n, w^n, \epsilon_n))_{n \in \mathbb{N}}$ is a monotonically decreasing sequence and

$$\lim_{n \to \infty} \|x^n - x^{n+1}\|_2^2 = 0.$$

Proof. Note that $\mathcal{J}(x^n, w^n, \epsilon_n) \ge \mathcal{J}(x^{n+1}, w^{n+1}, \epsilon_{n+1})$ for each $n = 1, 2, \dots,$ and

$$L = \mathcal{J}(x^1, w^0, \epsilon_0) \ge \mathcal{J}(x^n, w^n, \epsilon_n) \ge (w_j^n)^{-1}, \quad j = 1, \dots, N.$$

Hence, for each $n = 1, 2, \ldots$, the following estimates hold:

$$\begin{aligned}
&2[\mathcal{J}(x^n, w^n, \epsilon_n) - \mathcal{J}(x^{n+1}, w^{n+1}, \epsilon_{n+1})] \\
&\geq 2[\mathcal{J}(x^n, w^n, \epsilon_n) - \mathcal{J}(x^{n+1}, w^n, \epsilon_n)] = \langle x^n, x^n \rangle_{w^n} - \langle x^{n+1}, x^{n+1} \rangle_{w^n} \\
&= \langle x^n + x^{n+1}, x^n - x^{n+1} \rangle_{w^n} = \langle x^n - x^{n+1}, x^n - x^{n+1} \rangle_{w^n} \\
&= \sum_{j=1}^{N} w_j^n (x_j^n - x_j^{n+1})^2 \geq L^{-1} \|x^n - x^{n+1}\|_2^2.
\end{aligned}$$

In the third line, it is used that $\langle x^{n+1}, x^n - x^{n+1} \rangle_{w^n} = 0$ due to (49) since $x^n - x^{n+1}$ is contained in $\mathcal{N}$. ■

Moreover, if one assumes that $x^n \to \bar{x}$ and $\epsilon_n \to 0$, then, formally,

$$\mathcal{J}(x^n, w^n, \epsilon_n) \to \|\bar{x}\|_1.$$

Hence, one expects that this algorithm performs similar to ℓ_1-minimization. Indeed, the following convergence result holds.

Theorem 8. *Suppose $A \in \mathbb{R}^{m \times N}$ satisfies the NSP of order K with constant $\gamma < 1$. Use K in the update rule (52). Then, for each $y \in \mathbb{R}^m$, the sequence x^n produced by the algorithm converges to a vector $\bar{x}$, with $r_{K+1}(\bar{x}) = N \lim_{n \to \infty} \epsilon_n$, and the following holds:*

(i) If $\epsilon = \lim_{n \to \infty} \epsilon_n = 0$, then $\bar{x}$ is K-sparse; in this case, there is therefore a unique ℓ_1-minimizer x^*, and $\bar{x} = x^*$; moreover, one has, for $k \leq K$ and any $z \in \mathcal{F}(y)$,

$$\|z - \bar{x}\|_1 \leq \frac{2(1+\gamma)}{1-\gamma} \sigma_k(z)_1 \tag{56}$$

(ii) If $\epsilon = \lim_{n \to \infty} \epsilon_n > 0$, then $\bar{x} = x^\epsilon := \arg\min_{z \in \mathcal{F}(y)} \sum_{j=1}^{N} \left(z_j^2 + \epsilon^2 \right)^{1/2}$.

(iii) In this last case, if γ satisfies the stricter bound $\gamma < 1 - \frac{2}{K+2}$ (or, equivalently, if $\frac{2\gamma}{1-\gamma} < K$), then one has, for all $z \in \mathcal{F}(y)$ and any $k < K - \frac{2\gamma}{1-\gamma}$, that

$$\|z - \bar{x}\|_1 \leq \tilde{c} \sigma_k(z)_1, \qquad \text{with } \tilde{c} := \frac{2(1+\gamma)}{1-\gamma} \left[\frac{K - k + \frac{3}{2}}{K - k - \frac{2\gamma}{1-\gamma}} \right] \tag{57}$$

As a consequence, this case is excluded if $\mathcal{F}(y)$ contains a vector of sparsity $k < K - \frac{2\gamma}{1-\gamma}$.

Note that the approximation properties (56) and (57) are exactly of the same order as the one (6) provided by ℓ_1-minimization. However, in general, $\bar{x}$ is not necessarily an ℓ_1-minimizer, unless it coincides with a sparse solution. The proof of this result is not included and the interested reader is referred to [44, 69] for the details.

Rate of Convergence

It is instructive to show a further result concerning the local rate of convergence of this algorithm, which again uses the NSP as well as the optimality conditions introduced above. One assumes here that $\mathcal{F}(y)$ contains the k-sparse vector x^*. The algorithm produces a sequence x^n, which converges to x^*, as established above. One denotes the (unknown) support of the k-sparse vector x^* by T.

For now, one introduces an auxiliary sequence of error vectors $\eta^n \in \mathcal{N}$ via $\eta^n := x^n - x^*$ and

$$E_n := \|\eta^n\|_1 = \|x^* - x^n\|_1.$$

Theorem 8 guarantees that $E_n \to 0$ for $n \to \infty$. A useful technical result is reported next.

Lemma 4. *For any $z, z' \in \mathbb{R}^N$, and for any j,*

$$|\sigma_j(z)_1 - \sigma_j(z')_1| \le \|z - z'\|_1, \tag{58}$$

while for any $J > j$,

$$(J - j) r_J(z) \le \|z - z'\|_1 + \sigma_j(z')_1. \tag{59}$$

Proof. To prove (58), approximate z by a best j-term approximation $z'_{[j]} \in \Sigma_j$ of z' in ℓ_1. Then

$$\sigma_j(z)_1 \le \|z - z'_{[j]}\|_1 \le \|z - z'\|_1 + \sigma_j(z')_1,$$

and the result follows from symmetry. To prove (59), it suffices to note that $(J - j)\, r_J(z) \le \sigma_j(z)_1$. ∎

The following theorem gives a bound on the rate of convergence of E_n to zero.

Theorem 9. *Assume A satisfies the NSP of order K with constant γ. Suppose that $k < K - \frac{2\gamma}{1-\gamma}$, $0 < \rho < 1$, and $0 < \gamma < 1 - \frac{2}{K+2}$ are such that*

$$\mu := \frac{\gamma(1+\gamma)}{1-\rho}\left(1 + \frac{1}{K+1-k}\right) < 1.$$

Assume that $\mathcal{F}(y)$ contains a k-sparse vector x^* and let $T = \operatorname{supp}(x^*)$. Let n_0 be such that

$$E_{n_0} \le R^* := \rho \min_{i \in T} |x_i^*|. \tag{60}$$

Then, for all $n \ge n_0$, one has

$$E_{n+1} \le \mu\, E_n. \tag{61}$$

Consequently, x^n converges to x^* exponentially.

Proof. The relation (49) with $w = w^n$, $x^w = x^{n+1} = x^* + \eta^{n+1}$, and $\eta = x^{n+1} - x^* = \eta^{n+1}$ gives

$$\sum_{i=1}^{N} (x_i^* + \eta_i^{n+1}) \eta_i^{n+1} w_i^n = 0.$$

Rearranging the terms and using the fact that x^* is supported on T, one obtains

$$\sum_{i=1}^{N} |\eta_i^{n+1}|^2 w_i^n = -\sum_{i \in T} x_i^* \eta_i^{n+1} w_i^n = -\sum_{i \in T} \frac{x_i^*}{[(x_i^n)^2 + \epsilon_n^2]^{1/2}} \eta_i^{n+1}. \tag{62}$$

The proof of the theorem is by induction. Assume that $E_n \le R^*$ has already been established. Then, for all $i \in T$,

$$|\eta_i^n| \le \|\eta^n\|_1 = E_n \le \rho |x_i^*|,$$

so that

$$\frac{|x_i^*|}{[(x_i^n)^2 + \epsilon_n^2]^{1/2}} \le \frac{|x_i^*|}{|x_i^n|} = \frac{|x_i^*|}{|x_i^* + \eta_i^n|} \le \frac{1}{1-\rho} \tag{63}$$

and hence (62) combined with (63) and the NSP gives

$$\sum_{i=1}^{N} |\eta_i^{n+1}|^2 w_i^n \le \frac{1}{1-\rho} \|\eta_T^{n+1}\|_1 \le \frac{\gamma}{1-\rho} \|\eta_{T^c}^{n+1}\|_1$$

The Cauchy–Schwarz inequality combined with the above estimate yields

$$\begin{aligned} \|\eta_{T^c}^{n+1}\|_1^2 &\le \left(\sum_{i \in T^c} |\eta_i^{n+1}|^2 w_i^n \right) \left(\sum_{i \in T^c} [(x_i^n)^2 + \epsilon_n^2]^{1/2} \right) \\ &= \left(\sum_{i=1}^{N} |\eta_i^{n+1}|^2 w_i^n \right) \left(\sum_{i \in T^c} [(\eta_i^n)^2 + \epsilon_n^2]^{1/2} \right) \\ &\le \frac{\gamma}{1-\rho} \|\eta_{T^c}^{n+1}\|_1 \left(\|\eta^n\|_1 + N\epsilon_n \right). \end{aligned} \tag{64}$$

If $\eta^{n+1}_{T^c} = 0$, then $x^{n+1}_{T^c} = 0$. In this case x^{n+1} is k-sparse and the algorithm has stopped by definition; since $x^{n+1} - x^*$ is in the null space $\mathcal{N}$, which contains no k-sparse elements other than 0, one has already obtained the solution $x^{n+1} = x^*$. If $\eta^{n+1}_{T^c} \neq 0$, then canceling the factor $\|\eta^{n+1}_{T^c}\|_1$ in (64) yields

$$\|\eta^{n+1}_{T^c}\|_1 \leq \frac{\gamma}{1-\rho}\left(\|\eta^n\|_1 + N\epsilon_n\right),$$

and thus

$$\|\eta^{n+1}\|_1 = \|\eta^{n+1}_T\|_1 + \|\eta^{n+1}_{T^c}\|_1 \leq (1+\gamma)\|\eta^{n+1}_{T^c}\|_1 \leq \frac{\gamma(1+\gamma)}{1-\rho}\left(\|\eta^n\|_1 + N\epsilon_n\right). \tag{65}$$

Now, by (52) and (59), it follows

$$N\epsilon_n \leq r_{K+1}(x^n) \leq \frac{1}{K+1-k}(\|x^n - x^*\|_1 + \sigma_k(x^*)_1) = \frac{\|\eta^n\|_1}{K+1-k}, \tag{66}$$

since by assumption $\sigma_k(x^*)_1 = 0$. Together with (65), this yields the desired bound:

$$E_{n+1} = \|\eta^{n+1}\|_1 \leq \frac{\gamma(1+\gamma)}{1-\rho}\left(1 + \frac{1}{K+1-k}\right)\|\eta^n\|_1 = \mu E_n.$$

In particular, since $\mu < 1$, one has $E_{n+1} \leq R^*$, which completes the induction step. It follows that $E_{n+1} \leq \mu E_n$ for all $n \geq n_0$. ■

The linear rate (61) can be improved significantly, by a very simple modification of the rule of updating the weight:

$$w^{n+1}_j = \left((x^{n+1}_j)^2 + \epsilon^2_{n+1}\right)^{-\frac{2-\tau}{2}}, \quad j = 1, \ldots, N, \text{ for any } 0 < \tau < 1.$$

This corresponds to the substitution of the function $\mathcal{J}$ with

$$\mathcal{J}_\tau(z, w, \epsilon) := \frac{\tau}{2}\sum_{j=1}^N \left(z_j^2 w_j + \epsilon^2 w_j + \frac{2-\tau}{\tau}\frac{1}{w_j^{\frac{\tau}{2-\tau}}}\right),$$

where $z \in \mathbb{R}^N, w \in \mathbb{R}^N_+, \epsilon \in \mathbb{R}_+$. With this new up-to-date rule for the weight, which depends on $0 < \tau < 1$, one has formally, for $x^n \to \bar{x}$ and $\epsilon_n \to 0$,

$$\mathcal{J}_\tau(x^n, w^n, \epsilon_n) \to \|\bar{x}\|^\tau_\tau.$$

Hence such an iterative optimization tends to promote the ℓ_τ-quasi-norm minimization.

Surprisingly the rate of local convergence of this modified algorithm is superlinear; the rate is larger for smaller τ and approaches a quadratic rate as $\tau \to 0$. More precisely, the local error $E_n := \|x^n - x^*\|_\tau^\tau$ satisfies

$$E_{n+1} \leq \mu(\gamma, \tau) E_n^{2-\tau}, \tag{67}$$

where $\mu(\gamma, \tau) < 1$ for $\gamma > 0$ sufficiently small. The validity of (67) is restricted to x^n in a (small) ball centered at x^*. In particular, if x^0 is close enough to x^*, then (67) ensures the convergence of the algorithm to the k-sparse solution x^*; see Fig. 4.

Numerical Experiments

Figure 5 shows a typical *phase transition* diagram related to the (experimentally determined) probability of successful recovery of sparse vectors by means of the iteratively re-weighted least squares algorithm. For each point of this diagram with coordinates $(m/N, k/m) \in [0, 1]^2$, one shows the empirical success probability of recovery of a k-sparse vector $x \in \mathbb{R}^N$ from m measurements $y = Ax$. The brightness level corresponds to the probability. As measurement matrix a real random Fourier type matrix A was used, with entries given by

$$A_{k,j} = \cos(2\pi j \xi_k), \quad j = 1, \ldots, N,$$

and the $\xi_k, k = 1, \ldots, m$, are sampled independently and uniformly at random from $[0, 1]$. (Theorem 5 does not apply directly to real random Fourier matrices, but an analogous result concerning the RIP for such matrices can be found in [131].)

Figure 6 shows a section of a phase transition diagram related to the (experimentally determined) probability of successful recovery of sparse vectors from linear measurements $y = Ax$, where the matrix A has i.i.d. Gaussian entries. Here both m and N are fixed and only k is variable. This diagram establishes the transition from a situation of exact reconstruction for sparse vectors with high probability to very unlikely recovery for vectors with many nonzero entries. These numerical experiments used the iteratively re-weighted least squares algorithm with different parameters $0 < \tau \leq 1$. It is of interest to emphasize the enhanced success rate when using the algorithm for $\tau < 1$. Similarly, many other algorithms are tested by showing the corresponding phase transition diagrams and comparing them; see [12] for a detailed account of phase transitions for greedy algorithms and [49, 54] for ℓ_1-minimization.

This section is concluded by showing applications of ℓ_1-minimization methods to a real-life image recolorization problem [70, 71] in Fig. 7. The image is known completely only on very few colored portions, while on the remaining areas, only gray levels are provided. With this partial information, the use of ℓ_1-minimization with respect to wavelet or curvelet coefficients allows for high fidelity recolorization of the whole images.

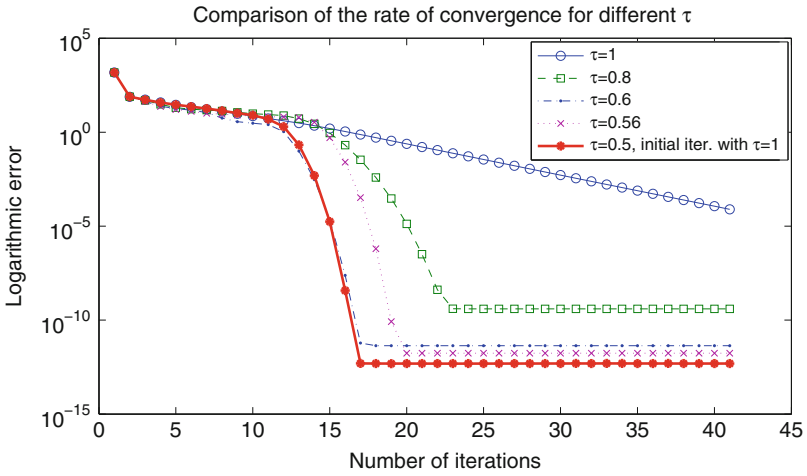


Fig. 4 The decay of logarithmic error is shown, as a function of the number of iterations of IRLS for different values of τ (1, 0.8, 0.6, 0.56). The results of an experiment are also shown, in which the initial 10 iterations are performed with $\tau = 1$ and the remaining iterations with $\tau = 0.5$

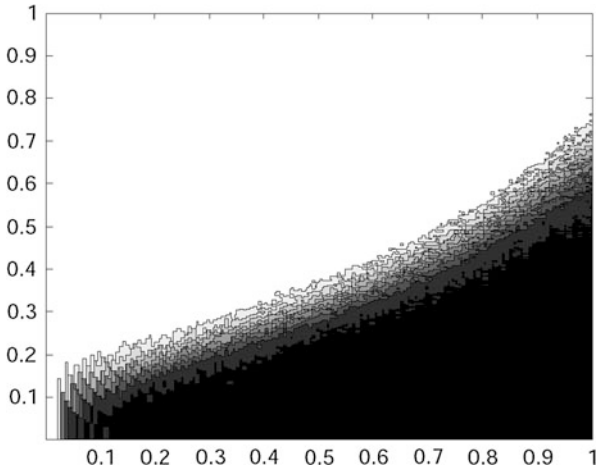


Fig. 5 Empirical success probability of recovery of k-sparse vectors $x \in \mathbb{R}^N$ from measurements $y = Ax$, where $A \in \mathbb{R}^{m \times N}$ is a real random Fourier matrix. The dimension $N = 300$ of the vectors is fixed. Each point of this diagram with coordinates $(m/N, k/m) \in [0, 1]^2$ indicates the empirical success probability of recovery, which is computed by running 100 experiments with randomly generated k-sparse vectors x and randomly generated matrix. The algorithm used for the recovery is the iteratively re-weighted least squares method tuned to promote ℓ_1-minimization

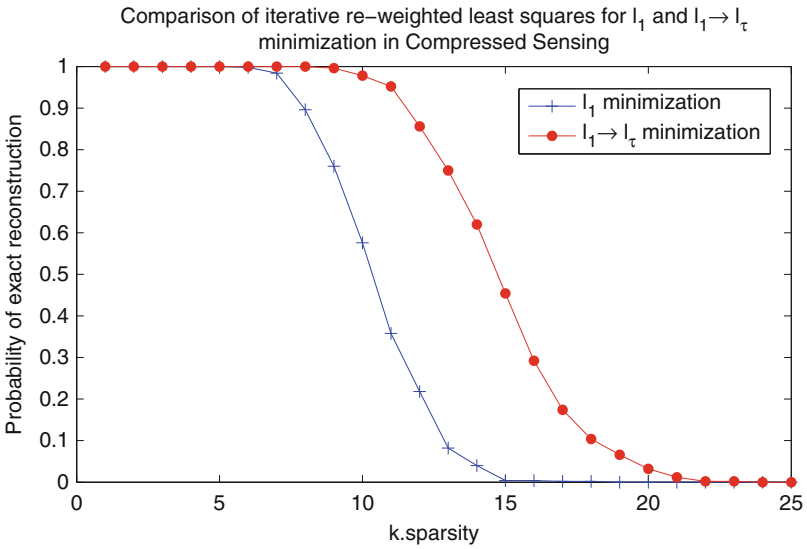


Fig. 6 Empirical success probability of recovery of a k-sparse vector $x \in \mathbb{R}^{250}$ from measurements $y = Ax$, where $A \in \mathbb{R}^{50\times 250}$ is Gaussian. The matrix is generated once; then, for each sparsity value k shown in the plot, 500 attempts were made, for randomly generated k-sparse vectors x. Two different IRLS algorithms were compared: one with weights inspired by ℓ_1-minimization and the IRLS with weights that gradually moved during the iterations from an ℓ_1- to an ℓ_τ-minimization goal, with final $\tau = 0.5$

Fig. 7 Iterations of the recolorization methods proposed in [70, 71] via ℓ_1 and total variation minimization, for the virtual restoration of the frescoes of A. Mantegna (1452), which were destroyed by a bombing during World War II. Only a few colored fragments of the images were saved from the disaster, together with good quality *gray*-level pictures dated to 1920

Extensions to Affine Low-Rank Minimization

Reformulated as a semidefinite program, the nuclear norm minimization (25) can be solved also by general methods from convex optimization [17]. Unfortunately, standard semidefinite programming tools work efficiently for solving nuclear norm minimization problems only for matrices up to size approximately 100×100.

Therefore, it is crucial to develop fast algorithms that are specialized to nuclear norm minimization (or other heuristics for rank minimization). Beside the already mentioned primal-dual algorithm of Sect. 4, several alternative approaches have been suggested so far; see, for instance, [20, 85, 97, 98, 104].

Borrowing a leaf from iteratively re-weighted least squares for ℓ_1-minimization, this section discusses an algorithm inspired by (25). Assume $X \in M_{n\times p}$, $0 \prec W = W^* \in M_{n\times n}$, and consider the functional,

$$\mathcal{J}(X,W) := \frac{1}{2}\left(\|W^{1/2}X\|_F^2 + \|W^{-1/2}\|_F^2\right), \tag{68}$$

where $\|\cdot\|_F$ is the Frobenius norm. In order to define the iteration below, one recalls that $\sigma_k(X)$ denotes the k th singular value of a matrix X.

IRLS-M algorithm for low-rank matrix recovery: Initialize by taking $W^0 := I \in M_{n\times n}$. Set $\varepsilon_0 := 1$, $K \geq k$, and $\gamma > 0$. Then recursively define, for $\ell = 1, 2, \ldots,$

$$X^\ell := \arg\min_{\mathcal{S}(X)=\mathcal{M}} \mathcal{J}(X, W^\ell)$$

and

$$\varepsilon_\ell := \min\left\{\varepsilon_{\ell-1}, \gamma\sigma_{K+1}(X^\ell)\right\}. \tag{69}$$

The update of the weight matrix W^ℓ follows the variational principle

$$W^\ell := \arg\min_{0\prec W=W^*\preceq\varepsilon_\ell^{-1}I} \mathcal{J}(X^\ell, W) \tag{70}$$

The algorithm stops if $\varepsilon_\ell = 0$; in this case, define $X^j := X^\ell$ for $j > \ell$. In general, the algorithm generates an infinite sequence $(X^\ell)_{\ell\in\mathbb{N}}$ of matrices.

The following convergence result can be shown with an analysis similar to the one done for the IRLS for sparse vector recovery, under the assumption that the measurement map $\mathcal{S}$ satisfies the restricted isometry property in Definition 4. One refers to [73] for details.

Proposition 2. *Consider the IRLS-M algorithm with parameters $\gamma = 1/n$ and $K \in \mathbb{N}$. Let $\mathcal{S} : M_{n\times p} \to \mathbb{C}^m$ be a surjective map with restricted isometry constants δ_{3K}, δ_{4K} satisfying $\eta = \frac{\sqrt{2}\delta_{4K}}{1-\delta_{3K}} < 1 - \frac{2}{K-2}$. Then, if there exists a k-rank matrix X satisfying $\mathcal{S}(X) = \mathcal{M}$ with $k < K - \frac{2\eta}{1-\eta}$, the sequence $(X^\ell)_{\ell\in\mathbb{N}}$ converges to X.*

Actually, one can prove something stronger: the IRLS-M algorithm is *robust*, in the sense that under the same conditions on the measurement map $\mathcal{S}$, the accumulation points of the IRLS-M algorithm are guaranteed to approximate an

arbitrary $X \in M_{n\times p}$ from the measurements $\mathscr{M} = \mathscr{S}(X)$ to within a factor of the *best k-rank approximation error* of X in the nuclear norm.

By combining both IRLS and IRLS-M, one can formulate an iteratively reweighted least squares to solve also mixed problems of the type (35).

5 Open Questions

The field of compressed sensing is rather young, so there remain many directions to be explored, and it is impossible to give an exhaustive list here. Below, we focus on two problems which seem to be rather hard and remained unsolved until the time of writing of this article.

Deterministic Compressed Sensing Matrices

So far, only several types of random matrices $A \in \mathbb{C}^{m\times N}$ are known to satisfy the RIP $\delta_s \le \delta \le 0.4$ (say) for

$$m = C_\delta s \log^\alpha(N) \tag{71}$$

for some constant C_δ and some exponent α (with high probability). This is a strong form of existence statement. It is open, however, to provide deterministic and explicit $m \times N$ matrices that satisfy the RIP $\delta_s \le \delta \le 0.5$ (say) in the desired range (71).

In order to show RIP estimates in the regime (71), one has to take into account cancellations of positive and negative (or more generally complex) entries in the matrix; see also Sect. 3. This is done "automatically" with probabilistic methods but seems to be much more difficult to exploit when the given matrix is deterministic. It may be conjectured that certain equiangular tight frames or the "Alltop matrix" in [123,143] do satisfy the RIP under (71). This is supported by numerical experiments in [123]. It is expected, however, that a proof is very hard and requires a good amount of analytic number theory. An involved deterministic construction that achieves the RIP in the regime $m \ge C s^{2-\varepsilon}$ – overcoming the quadratic bottleneck inherent to bounds via the coherence – was provided in [15, 16], but the best available estimates of ε are in the range of 10^{-26} [112]. One refers to [76, Chapter 6] for more background information on deterministic RIP matrices.

Another approach for deterministic constructions of CS matrices uses deterministic expander graphs [11]. Instead of the usual RIP, one shows that the adjacency matrix of such an expander graph has the 1-RIP, where the ℓ_2-norm is replaced by the ℓ_1-norm at each occurrence in (8). The 1-RIP also implies recovery by ℓ_1-minimization. The best known deterministic expanders [33] yield sparse recovery under the condition $m \ge C s (\log N)^{c \log^2(N)}$. Although the scaling in s is linear as desired, the term $(\log N)^{c \log^2(N)}$ grows faster than any polynomial in $\log N$.

Another drawback is that the deterministic expander graph is the output of a polynomial time algorithm, and it is questionable whether the resulting matrix can be regarded as *explicit*.

Removing Log-Factors in the Fourier-RIP Estimate

It is known [26, 76, 129, 131, 138] that a random partial Fourier matrix $A \in \mathbb{C}^{m\times N}$ satisfies the RIP with high probability provided

$$\frac{m}{\log(m)} \geq C_\delta s \log^2(s)\log(N).$$

(The condition stated in (19) implies this one.) It is conjectured that one can remove some of the log-factors. It must be hard, however, to improve this to a better estimate than $m \geq C_{\delta,\epsilon} s\log(N)\log(\log N)$. Indeed, this would imply an open conjecture of Talagrand [146] concerning the equivalence of the ℓ_1 and ℓ_2 norm of a linear combination of a subset of characters (complex exponentials). One refers to [76, Chapter 12.7] for more details.

Compressive Sensing with Nonlinear Measurements

As described in Sect. 3, the extension of compressed sensing towards nonlinear measurements is linked through the solution of problems of the type (35) which lead to semidefinite programming. Although some progress has been achieved, there is still the need to develop highly efficient algorithms for solving such problems in large dimensions. Besides these more practical aspects, also theoretical guarantees for algorithms to stably recover nearly sparse solutions consistent with given nonlinear measurements, for instance, phase retrieval problems beyond the case of Gaussian measurements, are a current field of very active research.

6 Conclusion

Compressive sensing established itself by now as a new sampling theory which exhibits fundamental and intriguing connections with several mathematical fields, such as probability, geometry of Banach spaces, harmonic analysis, theory of computability, and information-based complexity. The link to convex optimization and the development of very efficient and robust numerical methods make compressive sensing a concept useful for a broad spectrum of natural science and engineering applications, in particular, in signal and image processing and acquisition. It can be expected that compressive sensing will enter various branches of science and technology to notable effect.

New challenges are now emerging in numerical analysis and simulation where high-dimensional problems (e.g., stochastic partial differential equations in finance and electron structure calculations in chemistry and biochemistry) became the frontier. In this context, besides other forms of efficient approximation, such as sparse grid and tensor product methods [18], compressive sensing is a promising concept which is likely to cope with the "curse of dimensionality." In particular, further systematic developments of adaptivity in the presence of different scales, randomized algorithms, and an increasing role for combinatorial aspects of the underlying algorithms are examples of possible future developments, which are inspired by the successful history of compressive sensing [145].

Cross-References

Compressive sensing has connections with the following chapters of the book:

- ▸ Duality and Convex Programming
- ▸ Gabor Analysis for Imaging
- ▸ Large-Scale Inverse Problems in Imaging
- ▸ Linear Inverse Problems
- ▸ Mumford and Shah Model and its Applications to Image Segmentation and Image Restoration
- ▸ Numerical Methods and Applications in Total Variation Image Restoration
- ▸ Regularization Methods for Ill-Posed Problems
- ▸ Sampling Methods
- ▸ Splines and Multiresolution Analysis
- ▸ Starlet Transform in Astronomical Data Processing
- ▸ Supervised Learning by Support Vector Machines
- ▸ Synthetic Aperture Radar Imaging
- ▸ Tomography
- ▸ Total Variation in Imaging

Recommended Reading

The initial papers on the subject are [26, 29, 47]. The monograph [76] provides an introduction to compressive sensing. Further introductory articles can be found in [6, 21, 28, 61, 63, 69, 131, 136].

References

1. Achlioptas, D.: Database-friendly random projections. In: Proceedings of the 20th Annual ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, Santa Barbara, pp. 274–281 (2001)

2. Affentranger, F., Schneider, R.: Random projections of regular simplices. Discret. Comput. Geom. **7**(3), 219–226 (1992)
3. Ailon, N., Liberty, E.: Almost optimal unrestricted fast Johnson-Lindenstrauss transform. In: Symposium on Discrete Algorithms (SODA), San Francisco, (2011)
4. Alexeev, B., Bandeira, A.S., Fickus, M., Mixon, D.G.: Phase retrieval with polarization (2012). arXiv:1210.7752
5. Balan, R., Casazza, P., Edidin, D.: On signal reconstruction without phase. Appl. Comput. Harmon. Anal. **20**(3), 345–356 (2006)
6. Baraniuk, R.: Compressive sensing. IEEE Signal Process. Mag. **24**(4), 118–121 (2007)
7. Baraniuk, R.G., Davenport, M., DeVore, R.A., Wakin, M.: A simple proof of the restricted isometry property for random matrices. Constr. Approx. **28**(3), 253–263 (2008)
8. Bauschke, H.H., Combettes, P.-L., Luke, D.R.: Hybrid projection-reflection method for phase retrieval. J. Opt. Soc. Am. A **20**(6), 1025–1034 (2003)
9. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imaging Sci. **2**(1), 183–202 (2009)
10. Beck, A., Teboulle, M.: Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems. IEEE Trans. Image Process. **18**(11), 2419–2434 (2009)
11. Berinde, R., Gilbert, A.C., Indyk, P., Karloff, H., Strauss, M.: Combining geometry and combinatorics: a unified approach to sparse signal recovery. In: Proceedings of the 46th Annual Allerton Conference on Comunication, Control, and Computing 2008, Urbana, pp. 798–805. IEEE (2008)
12. Blanchard, J.D., Cartis, C., Tanner, J., Thompson, A.: Phase transitions for greedy sparse approximation algorithms. Appl. Comput. Harmon. Anal. **30**(2), 188–203 (2011)
13. Blumensath, T., Davies, M.: Iterative hard thresholding for compressed sensing. Appl. Comput. Harmon. Anal. **27**(3), 265–274 (2009)
14. Bobin, J., Starck, J.-L., Ottensamer, R.: Compressed sensing in astronomy. IEEE J. Sel. Top. Signal Process. **2**(5), 718–726 (2008)
15. Bourgain, J., Dilworth, S., Ford, K., Konyagin, S., Kutzarova, D.: Breaking the k^2-barrier for explicit RIP matrices. In: STOC'11, San Jose, pp. 637–644 (2011)
16. Bourgain, J., Dilworth, S., Ford, K., Konyagin, S., Kutzarova, D.: Explicit constructions of RIP matrices and related problems. Duke Math. J. **159**(1), 145–185 (2011)
17. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press, Cambridge/New York (2004)
18. Bungartz, H.-J., Griebel, M.: Sparse grids. Acta Numer. **13**, 147–269 (2004)
19. Cai, T., Zhang, A.: Sparse representation of a polytope and recovery of sparse signals and low-rank matrices. IEEE Trans. Inf. Theory **60**(1), 122–132 (2014)
20. Cai, J.-F., Candès, E.J., Shen, Z.: A singular value thresholding algorithm for matrix completion. SIAM J. Optim. **20**(4), 1956–1982 (2010)
21. Candès, E.J.: Compressive sampling. In: Proceedings of the International Congress of Mathematicians, Madrid (2006)
22. Candès, E.J.: The restricted isometry property and its implications for compressed sensing. C. R. Acad. Sci. Paris Ser. I Math. **346**, 589–592 (2008)
23. Candès, E.J., Li, X.: Solving quadratic equations via PhaseLift when there are about as many equations as unknowns. Found. Comput. Math. **14**(5), 1017–1026 (2014)
24. Candès, E.J., Plan, Y.: Tight oracle bounds for low-rank matrix recovery from a minimal number of random measurements. IEEE Trans. Inf. Theory **57**(4), 2342–2359 (2011)
25. Candès, E.J., Recht, B.: Exact matrix completion via convex optimization. Found. Comput. Math. **9**, 717–772 (2009)
26. Candès, E.J., Tao, T.: Near optimal signal recovery from random projections: universal encoding strategies? IEEE Trans. Inf. Theory **52**(12), 5406–5425 (2006)
27. Candès, E.J., Tao, T.: The power of convex relaxation: near-optimal matrix completion. IEEE Trans. Inf. Theory **56**(5), 2053–2080 (2010)
28. Candès, E., Wakin, M.: An introduction to compressive sampling. IEEE Signal Process. Mag. **25**(2), 21–30 (2008)

29. Candès, E.J., Tao, T., Romberg, J.: Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. IEEE Trans. Inf. Theory **52**(2), 489–509 (2006)
30. Candès, E.J., Romberg, J., Tao, T.: Stable signal recovery from incomplete and inaccurate measurements. Commun. Pure Appl. Math. **59**(8), 1207–1223 (2006)
31. Candès, E., Li, X., Soltanolkotabi, M.: Phase retrieval from masked Fourier transforms (2013, preprint)
32. Candès, E.J., Strohmer, T., Voroninski, V.: PhaseLift: exact and stable signal recovery from magnitude measurements via convex programming. Commun. Pure Appl. Math. **66**(8), 1241–1274 (2013)
33. Capalbo, M., Reingold, O., Vadhan, S., Wigderson, A.: Randomness conductors and constant-degree lossless expanders. In: Proceedings of the Thirty-Fourth Annual ACM Symposium on Theory of Computing, Montréal, pp. 659–668 (electronic). ACM (2002)
34. Chambolle, A., Pock, T.: A first-order primal-dual algorithm for convex problems with applications to imaging. J. Math. Imaging Vis. **40**, 120–145 (2011)
35. Chen, S.S., Donoho, D.L., Saunders, M.A.: Atomic decomposition by basis pursuit. SIAM J. Sci. Comput. **20**(1), 33–61 (1999)
36. Christensen, O.: An Introduction to Frames and Riesz Bases. Applied and Numerical Harmonic Analysis. Birkhäuser, Boston (2003)
37. Cline, A.K.: Rate of convergence of Lawson's algorithm. Math. Comput. **26**, 167–176 (1972)
38. Cohen, A., Dahmen, W., DeVore, R.A.: Compressed sensing and best k-term approximation. J. Am. Math. Soc. **22**(1), 211–231 (2009)
39. Combettes, P., Pesquet, J.-C.: A Douglas-Rachford splitting approach to nonsmooth convex variational signal recovery. IEEE J. Sel. Top. Signal Process. **1**(4), 564–574 (2007)
40. Combettes, P., Pesquet, J.-C.: Proximal splitting methods in signal processing. In: Bauschke, H., Burachik, R., Combettes, P., Elser, V., Luke, D., Wolkowicz, H. (eds.) Fixed-Point Algorithms for Inverse Problems in Science and Engineering, pp. 185–212. Springer, New York (2011)
41. Combettes, P., Wajs, V.: Signal recovery by proximal forward-backward splitting. Multisc. Model. Simul. **4**(4), 1168–1200 (electronic) (2005)
42. Cormode, G., Muthukrishnan, S.: Combinatorial algorithms for compressed sensing. In: CISS, Princeton (2006)
43. Daubechies, I., Defrise, M., De Mol, C.: An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. Commun. Pure Appl. Math. **57**(11), 1413–1457 (2004)
44. Daubechies, I., DeVore, R., Fornasier, M., Güntürk, C.: Iteratively re-weighted least squares minimization for sparse recovery. Commun. Pure Appl. Math. **63**(1), 1–38 (2010)
45. Davies, M., Gribonval, R.: Restricted isometry constants where ℓ^p sparse recovery can fail for $0 < p \leq 1$. IEEE Trans. Inf. Theory **55**(5), 2203–2214 (2009)
46. Do, B., Indyk, P., Price, E., Woodruff, D.: Lower bounds for sparse recovery. In: Proceedings of the SODA, Austin (2010)
47. Donoho, D.L.: Compressed sensing. IEEE Trans. Inf. Theory **52**(4), 1289–1306 (2006)
48. Donoho, D.L.: For most large underdetermined systems of linear equations the minimal l^1 solution is also the sparsest solution. Commun. Pure Appl. Anal. **59**(6), 797–829 (2006)
49. Donoho, D.L.: High-dimensional centrally symmetric polytopes with neighborliness proportional to dimension. Discret. Comput. Geom. **35**(4), 617–652 (2006)
50. Donoho, D.L., Elad, M.: Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ^1 minimization. Proc. Natl. Acad. Sci. USA **100**(5), 2197–2202 (2003)
51. Donoho, D.L., Huo, X.: Uncertainty principles and ideal atomic decompositions. IEEE Trans. Inf. Theory **47**(7), 2845–2862 (2001)
52. Donoho, D., Logan, B.: Signal recovery and the large sieve. SIAM J. Appl. Math. **52**(2), 577–591 (1992)
53. Donoho, D.L., Tanner, J.: Neighborliness of randomly projected simplices in high dimensions. Proc. Natl. Acad. Sci. USA **102**(27), 9452–9457 (2005)

54. Donoho, D.L., Tanner, J.: Counting faces of randomly-projected polytopes when the projection radically lowers dimension. J. Am. Math. Soc. **22**(1), 1–53 (2009)
55. Donoho, D.L., Tsaig, Y.: Fast solution of l1-norm minimization problems when the solution may be sparse. IEEE Trans. Inf. Theory **54**(11), 4789–4812 (2008)
56. Dorfman, R.: The detection of defective members of large populations. Ann. Stat. **14**, 436–440 (1943)
57. Douglas, J., Rachford, H.: On the numerical solution of heat conduction problems in two or three space variables. Trans. Am. Math. Soc. **82**, 421–439 (1956)
58. Duarte, M., Davenport, M., Takhar, D., Laska, J., Ting, S., Kelly, K., Baraniuk, R.: Single-pixel imaging via compressive sampling. IEEE Signal Process. Mag. **25**(2), 83–91 (2008)
59. Efron, B., Hastie, T., Johnstone, I., Tibshirani, R.: Least angle regression. Ann. Stat. **32**(2), 407–499 (2004)
60. Ehler, M., Fornasier, M., Sigl, J.: Quasi-linear compressed sensing. Multiscale Model. Simul. **12**(2), 725–754 (2014)
61. Elad, M.: Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing. Springer, New York (2010)
62. Elad, M., Bruckstein, A.M.: A generalized uncertainty principle and sparse representation in pairs of bases. IEEE Trans. Inf. Theory **48**(9), 2558–2567 (2002)
63. Eldar, Y., Kutyniok, G. (eds.): Compressed Sensing – Theory and Applications. Cambridge University Press, Cambridge/New York (2012)
64. Eldar, Y., Mendelson, S.: Phase retrieval: stability and recovery guarantees. Appl. Comput. Harmon. Anal. (to appear). doi:10.1016/j.acha.2013.08.003
65. Ender, J.: On compressive sensing applied to radar. Signal Process. **90**(5), 1402–1414 (2010)
66. Fannjiang, A., Yan, P., Strohmer, T.: Compressed remote sensing of sparse objects. SIAM J. Imaging Sci. **3**(3), 595–618 (2010)
67. Fazel, M.: Matrix rank minimization with applications. PhD thesis, Stanford University (2002)
68. Fienup, J.R.: Phase retrieval algorithms: a comparison. Appl. Opt. **21**(15), 2758–2769 (1982)
69. Fornasier, M.: Numerical methods for sparse recovery. In: Fornasier, M. (ed.) Theoretical Foundations and Numerical Methods for Sparse Recovery. Radon Series on Computational and Applied Mathematics, vol. 9, pp. 93–200. deGruyter, Berlin (2010). Papers based on the presentations of the summer school "Theoretical Foundations and Numerical Methods for Sparse Recovery", Vienna, Austria, 31 Aug-4 Sept 2009
70. Fornasier, M., March, R.: Restoration of color images by vector valued BV functions and variational calculus. SIAM J. Appl. Math. **68**(2), 437–460 (2007)
71. Fornasier, M., Ramlau, R., Teschke, G.: The application of joint sparsity and total variation minimization algorithms to a real-life art restoration problem. Adv. Comput. Math. **31**(1–3), 157–184 (2009)
72. Fornasier, M., Langer, A., Schönlieb, C.: A convergent overlapping domain decomposition method for total variation minimization. Numer. Math. **116**(4), 645–685 (2010)
73. Fornasier, M., Rauhut, H., Ward, R.: Low-rank matrix recovery via iteratively reweighted least squares minimization. SIAM J. Optim. **21**(4), 1614–1640 (2011)
74. Foucart, S.: A note on guaranteed sparse recovery via ℓ_1-minimization. Appl. Comput. Harmon. Anal. **29**(1), 97–103 (2010)
75. Foucart, S., Lai, M.: Sparsest solutions of underdetermined linear systems via ℓ_q-minimization for $0 < q \leq 1$. Appl. Comput. Harmon. Anal. **26**(3), 395–407 (2009)
76. Foucart, S., Rauhut, H.: A Mathematical Introduction to Compressive Sensing. Applied and Numerical Harmonic Analysis. Birkhäuser, Boston (2013)
77. Foucart, S., Pajor, A., Rauhut, H., Ullrich, T.: The Gelfand widths of ℓ_p-balls for $0 < p \leq 1$. J. Complex. **26**(6), 629–640 (2010)
78. Fuchs, J.J.: On sparse representations in arbitrary redundant bases. IEEE Trans. Inf. Theory **50**(6), 1341–1344 (2004)
79. Garnaev, A., Gluskin, E.: On widths of the Euclidean ball. Sov. Math. Dokl. **30**, 200–204 (1984)

80. Gilbert, A.C., Muthukrishnan, S., Guha, S., Indyk, P., Strauss, M.: Near-optimal sparse Fourier representations via sampling. In: Proceedings of the STOC'02, Montréal, pp. 152–161. Association for Computing Machinery (2002)
81. Gilbert, A.C., Muthukrishnan, S., Strauss, M.J.: Approximation of functions over redundant dictionaries using coherence. In: Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms, Baltimore, 12–14 Jan 2003, pp. 243–252. SIAM and Association for Computing Machinery (2003)
82. Gilbert, A.C., Strauss, M., Tropp, J.A., Vershynin, R.: One sketch for all: fast algorithms for compressed sensing. In: Proceedings of the 39th ACM Symposium Theory of Computing (STOC), San Diego (2007)
83. Glowinski, R., Le, T.: Augmented Lagrangian and Operator-Splitting Methods. SIAM, Philadelphia (1989)
84. Gluskin, E.: Norms of random matrices and widths of finite-dimensional sets. Math. USSR-Sb. **48**, 173–182 (1984)
85. Goldfarb, D., Ma, S.: Convergence of fixed point continuation algorithms for matrix rank minimization. Found. Comput. Math. **11**(2), 183–210 (2011)
86. Gorodnitsky, I., Rao, B.: Sparse signal reconstruction from limited data using FOCUSS: a re-weighted minimum norm algorithm. IEEE Trans. Signal Process. **45**(3), 600–616 (1997)
87. Gribonval, R., Nielsen, M.: Sparse representations in unions of bases. IEEE Trans. Inf. Theory **49**(12), 3320–3325 (2003)
88. Gross, D.: Recovering low-rank matrices from few coefficients in any basis. IEEE Trans. Inf. Theory **57**(3), 1548–1566 (2011)
89. Gross, D., Liu, Y.-K., Flammia, S.T., Becker, S., Eisert, J.: Quantum state tomography via compressed sensing. Phys. Rev. Lett. **105**, 150401 (2010)
90. Gross, D., Krahmer, F., Kueng, R.: Improved recovery guarantees for phase retrieval from coded diffraction patterns (2014, preprint)
91. Gross, D., Krahmer, F., Kueng, R.: A partial derandomization of PhaseLift using spherical designs. J. Fourier Anal. Appl. (to appear)
92. He, B., Yuan, X.: Convergence analysis of primal-dual algorithms for a saddle-point problem: from contraction perspective. SIAM J. Imaging Sci. **5**(1), 119–149 (2012)
93. Horn, R., Johnson, C.: Matrix Analysis. Cambridge University Press, Cambridge/New York (1990)
94. Hügel, M., Rauhut, H., Strohmer, T.: Remote sensing via l1-minimization. Found. Comput. Math. **14**, 115–150 (2014)
95. Johnson, W.B., Lindenstrauss, J. (eds.): Handbook of the Geometry of Banach Spaces, vol. I. North-Holland, Amsterdam (2001)
96. Kashin, B.: Diameters of some finite-dimensional sets and classes of smooth functions. Math. USSR Izv. **11**, 317–333 (1977)
97. Keshavan, R.H., Montanari, A., Oh, S.: Matrix completion from a few entries. IEEE Trans. Inf. Theory **56**, 2980–2998 (2010)
98. Keshavan, R.H., Montanari, A., Oh, S.: Matrix completion from noisy entries. J. Mach. Learn. Res. **11**, 2057–2078 (2010)
99. Krahmer, F., Rauhut, H.: Structured random measurements in signal processing. GAMM Mitteilungen. (to appear)
100. Krahmer, F., Ward, R.: New and improved Johnson-Lindenstrauss embeddings via the restricted isometry property. SIAM J. Math. Anal. **43**(3), 1269–1281 (2011)
101. Krahmer, F., Mendelson, S., Rauhut, H.: Suprema of chaos processes and the restricted isometry property. Commun. Pure Appl. Math. (to appear). doi:10.1002/cpa.21504
102. Lawson, C.: Contributions to the theory of linear least maximum approximation. PhD thesis, University of California, Los Angeles (1961)
103. Ledoux, M., Talagrand, M.: Probability in Banach Spaces. Springer, Berlin/New York (1991)
104. Lee, K., Bresler, Y.: ADMiRA: atomic decomposition for minimum rank approximation. IEEE Trans. Inf. Theory **56**(9), 4402–4416 (2010)

105. Li, X., Voroninski, V.: Sparse signal recovery from quadratic measurements via convex programming (2013). arXiv:1209.4785
106. Logan, B.: Properties of high-pass signals. PhD thesis, Columbia University (1965)
107. Lorentz, G.G., von Golitschek, M., Makovoz, Y.: Constructive Approximation: Advanced Problems. Springer, Berlin (1996)
108. Mallat, S.G., Zhang, Z.: Matching pursuits with time-frequency dictionaries. IEEE Trans. Signal Process. **41**(12), 3397–3415 (1993)
109. Marple, S.: Digital Spectral Analysis with Applications. Prentice-Hall, Englewood Cliffs (1987)
110. Mendelson, S., Pajor, A., Tomczak Jaegermann, N.: Uniform uncertainty principle for Bernoulli and subgaussian ensembles. Constr. Approx. **28**(3), 277–289 (2009)
111. Millane, R.: Phase retrieval in crystallography and optics. J. Opt. Soc. Am. A **7**(3), 394–411 (1990)
112. Mixon, D.: Short, fat matrices. Blog (2013)
113. Mohan, K., Fazel, M.: Reweighted nuclear norm minimization with application to system identification. In: Proceedings of the American Control Conference, Baltimore, pp. 2953–2959 (2010)
114. Natarajan, B.K.: Sparse approximate solutions to linear systems. SIAM J. Comput. **24**, 227–234 (1995)
115. Nesterov, Y., Nemirovskii, A.: Interior-Point Polynomial Algorithms in Convex Programming. Volume 13 of SIAM Studies in Applied Mathematics. Society for Industrial and Applied Mathematics (SIAM), Philadelphia (1994)
116. Netrapalli, P., Jain, P., Sanghavi, S.: Phase retrieval using alternating minimization (2013). arXiv:1306.0160
117. Novak, E.: Optimal recovery and n-widths for convex classes of functions. J. Approx. Theory **80**(3), 390–408 (1995)
118. Ohlsson, H., Yang, A.Y., Dong, R., Sastry, S.S.: Nonlinear basis pursuit. In: 47th Asilomar Conference on Signals, Systems and Computers, Pacific Grove (2013)
119. Osborne, M., Presnell, B., Turlach, B.: A new approach to variable selection in least squares problems. IMA J. Numer. Anal. **20**(3), 389–403 (2000)
120. Osborne, M., Presnell, B., Turlach, B.: On the LASSO and its dual. J. Comput. Graph. Stat. **9**(2), 319–337 (2000)
121. Oymak, S., Mohan, K., Fazel, M., Hassibi, B.: A simplified approach to recovery conditions for low-rank matrices. In: Proceedings of the IEEE International Symposium on Information Theory (ISIT), St. Petersburg (2011)
122. Pfander, G.E., Rauhut, H.: Sparsity in time-frequency representations. J. Fourier Anal. Appl. **16**(2), 233–260 (2010)
123. Pfander, G.E., Rauhut, H., Tanner, J.: Identification of matrices having a sparse representation. IEEE Trans. Signal Process. **56**(11), 5376–5388 (2008)
124. Pfander, G.E., Rauhut, H., Tropp, J.A.: The restricted isometry property for time-frequency structured random matrices. Probab. Theory Relat. Fields **156**, 707–737 (2013)
125. Pock, T., Chambolle, A.: Diagonal preconditioning for first order primal-dual algorithms in convex optimization. In: IEEE International Conference Computer Vision (ICCV), Barcelona, pp. 1762–1769 (2011)
126. Pock, T., Cremers, D., Bischof, H., Chambolle, A.: An algorithm for minimizing the Mumford-Shah functional. In: ICCV Proceedings, Kyoto. Springer (2009)
127. Prony, R.: Essai expérimental et analytique sur les lois de la Dilatabilité des uides élastique et sur celles de la Force expansive de la vapeur de lÃoeau et de la vapeur de l'alkool, à différentes températures. J. École Polytechnique **1**, 24–76 (1795)
128. Rauhut, H.: Random sampling of sparse trigonometric polynomials. Appl. Comput. Harmon. Anal. **22**(1), 16–42 (2007)
129. Rauhut, H.: Stability results for random sampling of sparse trigonometric polynomials. IEEE Trans. Inf Theory **54**(12), 5661–5670 (2008)

130. Rauhut, H.: Circulant and Toeplitz matrices in compressed sensing. In: Proceedings of the SPARS'09 (2009)
131. Rauhut, H.: Compressive sensing and structured random matrices. In: Fornasier, M. (ed.) Theoretical Foundations and Numerical Methods for Sparse Recovery. Radon Series on Computational and Applied Mathematics, vol. 9, pp. 1–92. deGruyter, Berlin (2010). Papers based on the presentations of the summer school "Theoretical Foundations and Numerical Methods for Sparse Recovery", Vienna, Austria, 31 Aug-4 Sept 2009
132. Rauhut, H., Ward, R.: Interpolation via weighted l1 minimization (2013). ArXiv:1308.0759
133. Rauhut, H., Romberg, J.K., Tropp, J.A.: Restricted isometries for partial random circulant matrices. Appl. Comput. Harmon. Anal. **32**(2), 242–254 (2012)
134. Recht, B.: A simpler approach to matrix completion. J. Mach. Learn. Res. **12**, 3413–3430 (2012)
135. Recht, B., Fazel, M., Parrilo, P.: Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. SIAM Rev. **52**(3), 471–501 (2010)
136. Romberg, J.: Imaging via compressive sampling. IEEE Signal Process. Mag. **25**(2), 14–20 (2008)
137. Romberg, J.K.: Compressive sensing by random convolution. SIAM J. Imaging Sci. **2**(4), 1098–1128 (2009)
138. Rudelson, M., Vershynin, R.: On sparse reconstruction from Fourier and Gaussian measurements. Commun. Pure Appl. Math. **61**, 1025–1045 (2008)
139. Rudin, L., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. Physica D **60**(1–4), 259–268 (1992)
140. Santosa, F., Symes, W.: Linear inversion of band-limited reflection seismograms. SIAM J. Sci. Stat. Comput. **7**(4), 1307–1330 (1986)
141. Schnass, K., Vandergheynst, P.: Dictionary preconditioning for greedy algorithms. IEEE Trans. Signal Process. **56**(5), 1994–2002 (2008)
142. Starck, J.-L., Murtagh, F., Fadili, J.: Sparse Image and Signal Processing Wavelets, Curvelets, Morphological Diversity, xvii, p. 316. Cambridge University Press, Cambridge (2010)
143. Strohmer, T., Heath, R.W., Jr.: Grassmannian frames with applications to coding and communication. Appl. Comput. Harmon. Anal. **14**(3), 257–275 (2003)
144. Strohmer, T., Hermann, M.: Compressed sensing radar. In: IEEE Proceedings of the International Conference on Acoustic, Speech, and Signal Processing, Las Vegas, pp. 1509–1512 (2008)
145. Tadmor, E.: Numerical methods for nonlinear partial differential equations. In: Meyers, R.A. (ed.) Encyclopedia of Complexity and Systems Science. Springer, New York/London (2009)
146. Talagrand, M.: Selecting a proportion of characters. Isr. J. Math. **108**, 173–191 (1998)
147. Tauböck, G., Hlawatsch, F., Eiwen, D., Rauhut, H.: Compressive estimation of doubly selective channels in multicarrier systems: leakage effects and sparsity-enhancing processing. IEEE J. Sel. Top. Signal Process. **4**(2), 255–271 (2010)
148. Taylor, H., Banks, S., McCoy, J.: Deconvolution with the ℓ_1-norm. Geophys. J. Int. **44**(1), 39–52 (1979)
149. Tibshirani, R.: Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B **58**(1), 267–288 (1996)
150. Traub, J., Wasilkowski, G., Wo'zniakowski, H.: Information-Based Complexity. Computer Science and Scientific Computing. Academic, Boston (1988)
151. Tropp, J.A.: Greed is good: algorithmic results for sparse approximation. IEEE Trans. Inf. Theory **50**(10), 2231–2242 (2004)
152. Tropp, J.A.: Just relax: convex programming methods for identifying sparse signals in noise. IEEE Trans. Inf. Theory **51**(3), 1030–1051 (2006)
153. Tropp, J., Needell, D.: CoSaMP: iterative signal recovery from incomplete and inaccurate samples. Appl. Comput. Harmon. Anal. **26**(3), 301–321 (2009)
154. Tropp, J.A., Laska, J.N., Duarte, M.F., Romberg, J.K., Baraniuk, R.G.: Beyond nyquist: efficient sampling of sparse bandlimited signals. IEEE Trans. Inf. Theory **56**(1), 520–544 (2010)

155. Unser, M.: Sampling—50 years after Shannon. Proc. IEEE **88**(4), 569–587 (2000)
156. van den Berg, E., Friedlander, M.: Probing the Pareto frontier for basis pursuit solutions. SIAM J. Sci. Comput. **31**(2), 890–912 (2008)
157. Vybiral, J.: Widths of embeddings in function spaces. J. Complex. **24**(4), 545–570 (2008)
158. Wagner, G., Schmieder, P., Stern, A., Hoch, J.: Application of non-linear sampling schemes to cosy-type spectra. J. Biomol. NMR **3**(5), 569 (1993)
159. Willett, R., Marcia, R., Nichols, J.: Compressed sensing for practical optical imaging systems: a tutorial. Opt. Eng. **50**(7), 072601–072601–13 (2011)
160. Willett, R., Duarte, M., Davenport, M., Baraniuk, R.: Sparsity and structure in hyperspectral imaging: sensing, reconstruction, and target detection. IEEE Signal Proc. Mag. **31**(1), 116–126 (2014)

Duality and Convex Programming

Jonathan M. Borwein and D. Russell Luke

Contents

1 Introduction 258
Linear Inverse Problems with Convex Constraints 262
Imaging with Missing Data 263
Image Denoising and Deconvolution 266
Inverse Scattering 268
Fredholm Integral Equations 269
2 Background 271
Lipschitzian Properties 272
Subdifferentials 274
3 Duality and Convex Analysis 281
Fenchel Conjugation 281
Fenchel Duality 284
Applications 286
Optimality and Lagrange Multipliers 289
Variational Principles 291
Fixed Point Theory and Monotone Operators 292
4 Case Studies 293
Linear Inverse Problems with Convex Constraints 293
Imaging with Missing Data 295
Inverse Scattering 296
Fredholm Integral Equations 297
5 Open Questions 298
6 Conclusion 298

J.M. Borwein (✉)
School of Mathematical and Physical Sciences, University of Newcastle, Newcastle, NSW, Australia
e-mail: Jonathan.Borwein@newcastle.edu.au; jon.borwein@gmail.com

D.R. Luke
Institute of Numerical and Applied Mathematics, Georg-August-Universität Göttingen, Göttingen, Germany
e-mail: r.luke@math.uni-goettingen.de

© Springer Science+Business Media New York 2015

O. Scherzer (ed.), *Handbook of Mathematical Methods in Imaging*,
DOI 10.1007/978-1-4939-0790-8_7

Cross-References 299
References 299

Abstract

This chapter surveys key concepts in convex duality theory and their application to the analysis and numerical solution of problem archetypes in imaging.

Convex analysis, Variational analysis, Duality

1 Introduction

An image is worth a thousand words, the joke goes, but takes up a million times the memory – all the more reason to be efficient when computing with images. Whether one is determining a "best" image according to some criteria or applying a "fast" iterative algorithm for processing images, the theory of optimization and variational analysis lies at the heart of achieving these goals. Success or failure hinges on the abstract structure of the task at hand. For many practitioners, these details are secondary: if the picture looks good, it *is* good. For the specialist in variational analysis and optimization, however, it is what is went into constructing the image that matters: if it is *convex*, it is good.

This chapter surveys more than a half-a-century of work in convex analysis that has played a fundamental role in the development of computational imaging and brings to light as many of the contributors to this field as possible. There is no shortage of good books on convex and variational analysis; interested readers are referred to the modern references [4, 6, 24, 27, 31, 49, 61, 71, 72, 80, 93, 99, 103, 106, 107, 110, 111, 118]. References focused more on imaging and signal processing, but with a decidedly variational flavor, include [3, 39, 109]. For general references on numerical optimization, see [21, 35, 42, 86, 97, 117].

For many years, the dominant distinction in applied mathematics between problem types has rested upon linearity, or lack thereof. This categorization still holds sway today with nonlinear problems largely regarded as "hard," while linear problems are generally considered "easy." But since the advent of the *interior point revolution* [96], at least in linear optimization, it is more or less agreed that *nonconvexity*, not nonlinearity, more accurately delineates hard from easy. The goal of this chapter is to make this case more broad. Indeed, for convex sets topological, algebraic, and geometric notions often coincide, and so the tools of convex analysis provide not only for a tremendous synthesis of ideas but also for key insights, whose dividends are efficient algorithms for solving large (*infinite dimensional*) problems, and indeed even large *nonlinear* problems.

This chapter concerns different instances of a single optimization model. This model accounts for the vast majority of variational problems appearing in imaging science:

$$\begin{array}{ll} \underset{x \in C \subset X}{\text{minimize}} & I_\varphi(x) \\ \text{subject to} & Ax \in D. \end{array} \tag{1}$$

Here, X and Y are real Banach spaces with continuous duals X^* and Y^*, C and D are closed and convex, $A : X \to Y$ is a continuous linear operator, and the integral functional $I_\varphi(x) := \int_T \varphi(x(t))\mu(dt)$ is defined on some vector subspace $L_p(T, \mu)$ of X for μ, a complete totally finite measure on some measure space T. The integral operator I_φ is an *entropy* with integrand $\varphi : \mathbb{R} \to]-\infty, +\infty]$ a closed convex function. This provides an extremely flexible framework that specializes to most of the instances of interest and is general enough to extend results to non-Hilbert space settings. The most common examples are

$$\text{Burg entropy:} \quad \varphi(x) := -\ln(x) \tag{2}$$

$$\text{Shannon–Boltzmann entropy:} \quad \varphi(x) := x \ln(x) \tag{3}$$

$$\text{Fermi–Dirac entropy :} \quad \varphi(x) := x \ln(x) + (1 - x) \ln(1 - x) \tag{4}$$

$$L_p \text{ norm} \quad \varphi(x) := \frac{\|x\|^p}{p} \tag{5}$$

$$L_p \text{ entropy} \quad \varphi(x) := \begin{cases} \frac{x^p}{p} & x \geq 0 \\ +\infty & \text{else} \end{cases} \tag{6}$$

$$\text{Total variation} \quad \varphi(x) := |\nabla x|. \tag{7}$$

See [18, 25, 26, 34, 37, 43, 44, 56, 64, 112] for these and other entropies.

There is a rich correspondence between the algorithmic approach to applications implicit in the variational formulation (1) and the prevalent *feasibility* approach to problems. Here, one considers the problem of finding the point x that lies in the intersection of the constraint sets:

$$\text{find } x \in C \cap S \quad \text{where} \quad S := \{x \in X \mid Ax \in D\}.$$

In the case where the intersection is quite large, one might wish to find the point in the intersection in some sense closest to a reference point x_0 (frequently the origin). It is the job of the objective in (1) to pick the element of $C \cap S$ that has the desired properties, that is, to pick the *best approximation*. The feasibility formulation suggests very naturally projection algorithms for finding the intersection whereby one applies the constraints one at a time in some fashion, e.g., cyclically, or at random [5, 30, 42, 51, 52, 119]. This is quite a powerful framework as it provides a great deal of flexibility and is amenable to parallelization for large-scale problems. Many of the algorithms for feasibility problems have counterparts for the more general best approximation problems [6, 9, 10, 58, 88]. For studies of these algorithms in nonconvex settings, see [2, 7, 8, 11–14, 29, 38, 53, 68, 78, 79, 87–89]. The projection

algorithms that are central to convex feasibility and best approximation problems play a key role in algorithms for solving the problems considered here.

Before detailing specific applications, it is useful to state a general duality result for problem (1) that motivates the convex analytic approach. One of the more central tools is the *Fenchel conjugate* [62] of a mapping $f : X \to [-\infty, +\infty]$, denoted $f^* : X^* \to [-\infty, +\infty]$ and defined by

$$f^*(x^*) = \sup_{x \in X} \{\langle x^*, x\rangle - f(x)\}.$$

The conjugate is always convex (as a supremum of affine functions), while $f = f^{**}|_X$ exactly if f is convex, proper (not everywhere infinite), and lower semi-continuous (lsc) [24,61]. Here and below, unless otherwise specified, X is a normed space with dual X^*. The following theorem uses constraint qualifications involving the concept of the *core* of a set, the *effective domain* of a function (dom f), and the points of continuity of a function (cont f).

Definition 1 (Core). The core of a set $F \subset X$ is defined by $x \in \text{core } F$ if for each $h \in \{x \in X \mid \|x\| = 1\}$, there exists $\delta > 0$ so that $x + th \in F$ for all $0 \leq t \leq \delta$.

It is clear from the definition that int $F \subset$ core F. If F is a convex subset of a Euclidean space, or if F is closed, then the core and the interior are *identical* [27, Theorem 4.1.4].

Theorem 1 (Fenchel Duality [24, Theorems 2.3.4 and 4.4.18]). *Let X and Y be Banach spaces, let $f : X \to (-\infty, +\infty]$ and $g : Y \to (-\infty, +\infty]$ and let $A : X \to Y$ be a bounded linear map. Define the primal and dual values p, $d \in [-\infty, +\infty]$ by the Fenchel problems*

$$\begin{aligned} p &= \inf_{x \in X} \{f(x) + g(Ax)\} \\ d &= \sup_{y^* \in Y^*} \{-f^*(A^*y^*) - g^*(-y^*)\}. \end{aligned} \tag{8}$$

Then these values satisfy the weak duality inequality $p \geq d$.

If f, g are convex and satisfy either

$$0 \in \text{core } (\text{dom } g - A \,\text{dom } f) \quad \textit{with } f \textit{ and } g \textit{ lsc}, \tag{9}$$

or

$$A \,\text{dom } f \cap \text{cont } g \neq \emptyset, \tag{10}$$

then $p = d$, and the supremum to the dual problem is attained if finite.

Applying Theorem 1 to problem (1) yields $f(x) = I_\varphi(x) + \iota_C(x)$ and $g(y) = \iota_D(y)$ where ι_F is the *indicator function* of the set F:

$$\iota_F(x) := \begin{cases} 0 & \text{if } x \in F \\ +\infty & \text{else.} \end{cases} \tag{11}$$

The tools of convex analysis and the phenomenon of duality are central to formulating, analyzing, and solving application problems. Already apparent from the general application above is the necessity for a calculus of Fenchel conjugation in order to compute the conjugate of sums of functions. In some specific cases, one can arrive at the same conclusion with less theoretical overhead, but this is at the cost of missing out more general structures that are not necessarily automatic in other settings.

Duality has a long-established place in economics where primal and dual problems have direct interpretations in the context of the theory of zero-sum games, or where Lagrange multipliers and dual variables are understood, for instance, as shadow prices. In imaging, there is not as often an easy interplay between the physical interpretation of primal and dual problems. Duality has been used toward a variety of ends in contemporary image and signal processing, the majority of them, however, having to do with algorithms [17, 33, 34, 43–46, 54, 55, 57, 70, 73, 90, 116]. Nevertheless, the dual perspective yields new statistical or information theoretic insight into image processing problems, in addition to faster algorithms. Since the publication of the first edition of this handbook, interest in convex duality theory has only continued to grow. This is, in part, due to the now ubiquitous application of convex duality techniques to nonconvex problems; as heuristics, through appropriate convex relaxations, or otherwise [48, 114, 115]. As a measure of this growing interest, a *Web of Science* search for articles published between 2011 and 2013 having either "convex relaxation" or "primal dual" in the title, abstract, or keywords, returns a combined count of approximately 1,500 articles.

Modern optimization packages heavily exploit duality and convex analysis. A new trend that has matured in recent years is the field of *computational* convex analysis which employs symbolic, numerical, and hybrid computations of objects like the Fenchel conjugate [23, 65, 81–85, 91, 98]. Software suites that rely heavily on a convex duality approach include: *CCA* (Computational Convex Analysis, http://atoms.scilab.org/toolboxes/CCA) for *Scilab*, *CVX* and its extensions (http://cvxr.com/cvx/) for *MATLAB* (including a *C* code generator) [65, 91], and *S-CAT* (Symbolic Convex Analysis Toolkit, http://web.cs.dal.ca/~chamilto/research.html) for *Maple* [23]. For a review of the computational aspects of convex analysis see [84].

In this chapter, five main applications illustrate the variational analytical approach to problem solving: linear inverse problems with convex constraints, compressive imaging, image denoising and deconvolution, nonlinear inverse scattering, and finally Fredholm integral equations. A brief review of these applications is presented below. Subsequent sections develop the tools for their

analysis. At the end of the chapter these applications are revisited in light of the convex analytical tools collected along the way.

The application of compressive sensing, or more generally sparsity optimization leads to a new theme that is emerging out of the theory of convex relaxations for nonconvex problems, namely direct global *nonconvex* methods for structured *nonconvex* optimization problems. Starting with the seminal papers of Candés and Tao [40, 41], the theory of convex relaxation for finding the sparsest vector satisfying an underdetermined affine constraint has concentrated on determining sufficient conditions under which the solution to a relaxation of the problem to ℓ_1 minimization with an affine equality constraint corresponds exactly to the global solution of the original nonconvex sparsity optimization problem. These conditions have been used in recent years to guarantee global convergence of simple projected gradient- and alternating projections-type algorithms for solving simpler nonconvex optimization problems whose global solutions correspond to a solution of the original problem (see [15, 16, 19, 20, 63, 69] and references therein). This is the natural point at which this chapter leaves off, and the frontier of *nonconvex* programming begins.

Linear Inverse Problems with Convex Constraints

Let X be a Hilbert space and $\varphi(x) := \frac{1}{2}\|x\|^2$. The integral functional I_φ is the usual L_2 norm and the solution is the closest feasible point to the origin:

$$\begin{array}{ll} \underset{x\in C\subset X}{\text{minimize}} & \frac{1}{2}\|x\|^2 \\ \text{subject to} & Ax = b. \end{array} \tag{12}$$

Levi, for instance, used this variational formulation to determine the minimum energy band-limited signal that matched N measurements $b \in \mathbb{R}^n$ with the model $A : X \to \mathbb{R}^n$ [77]. Note that the signal space is infinite dimensional while the measurement space is finite dimensional, a common situation in practice. Potter and Arun [100] recognized a much broader applicability of this variational formulation to remote sensing and medical imaging and applied duality theory to characterize solutions to (12) by $\overline{x} = P_C A^*(\overline{y})$, where $\overline{y} \in Y$ satisfies $b = AP_C A^*\overline{y}$ [100, Theorem 1]. Particularly attractive is the feature that when Y is finite dimensional, these formulas yield a finite dimensional approach to an infinite dimensional problem. The numerical algorithm suggested by Potter and Arun is an iterative procedure in the dual variables:

$$y_{j+1} = y_j + \gamma(b - AP_C A^* y_j) \quad j = 0, 1, 2, \ldots \tag{13}$$

The optimality condition and numerical algorithms are explored at the end of this chapter.

As satisfying as this theory is, there is a crucial assumption in the theorem of Potter and Arun about the existence of $\overline{y} \in Y$ such that $b = AP_C A^* \overline{y}$; one need only to consider linear least squares, for an example, where the primal problem is well posed, but no such $\overline{y}$ exists [22]. A specialization of Theorem 1 to the case of linear constraints facilitates the argument. The next corollary is a specialization of Theorem 1, where g is the indicator function of the point b in the linear constraint.

Corollary 1 (Fenchel Duality for Linear Constraints). *Given any $f : X \to (-\infty, \infty]$, any bounded linear map $A : X \to Y$, and any element $b \in Y$, the following weak duality inequality holds:*

$$\inf_{x \in X} \{f(x) \mid Ax = b\} \geq \sup_{y^* \in Y^*} \{\langle b,\ y^* \rangle - f^*(A^* y^*)\}.$$

If f is lsc and convex and $b \in \operatorname{core}(A \operatorname{dom} f)$, then equality holds and the supremum is attained if finite.

Suppose that $C = X$, a Hilbert space and $A : X \to X$. The Fenchel dual to (12) is

$$\underset{y \in X}{\text{maximize}}\ \langle y,\ b \rangle - \frac{1}{2} \|A^* y\|^2. \tag{14}$$

(The L_2 norm is self-dual.) Suppose that the primal problem (12) is *feasible*, that is, $b \in \operatorname{range}(A)$. The objective in (14) is convex and differentiable, so elementary calculus (Fermat's rule) yields the optimal solution $\overline{y}$ with $AA^* \overline{y} = b$, assuming $\overline{y}$ exists. If the range of A is strictly larger than that of AA^*, however, it is possible to have $b \in \operatorname{range}(A)$ but $b \notin \operatorname{range}(AA^*)$, in which case the optimal solution $\overline{x}$ to (12) is *not* equal to $A^* \overline{y}$, since $\overline{y}$ is not attained. For a concrete example see [22, Example 2.1].

Imaging with Missing Data

Let $X = \mathbb{R}^n$ and $\varphi(x) := \|x\|_p$ for $p = 0$ or $p = 1$. The case $p = 1$ is the ℓ_1 norm, and by $\|x\|_0$ is meant the function

$$\|x\|_0 := \sum_j |\operatorname{sign}(x_j)|,$$

where $\operatorname{sign}(0) := 0$. This yields the optimization problem

$$\begin{array}{ll} \underset{x \in \mathbb{R}^n}{\text{minimize}} & \|x\|_p \\ \text{subject to} & Ax = b. \end{array} \tag{15}$$

This model has received a great deal of attention recently in applications of compressive sensing where the number of measurements is much smaller than the dimension of the signal space, that is, $b \in \mathbb{R}^m$ for $m \ll n$. This problem is well known in statistics as the missing data problem.

For ℓ_1 optimization ($p = 1$), the seminal work of Candés and Tao establishes probabilistic criteria for when the solution to (15) is unique and exactly matches the true signal x_* [41]. Sparsity of the original signal x_* and the algebraic structure of the matrix A are key requirements. Convex analysis easily yields a geometric interpretation of these facts. The dual to this problem is the linear program

$$\begin{array}{ll} \underset{y\in\mathbb{R}^m}{\text{maximize}} & b^T y \\ \text{subject to} & (A^* y)_j \in [-1,1] \quad j = 1,2,\ldots,n. \end{array} \tag{16}$$

Deriving this dual is one of the goals of this chapter. Elementary facts from linear programming guarantee that the solution includes a vertex of the polyhedron described by the constraints, and hence, assuming A is full rank, there can be at most m active constraints. The number of active constraints in the dual problem provides an upper bound on the number of nonzero elements in the primal variable – the signal to be recovered. Unless the number of nonzero elements of x_* is less than the number of measurements m, there is no hope of uniquely recovering x_*. The uniqueness of solutions to the primal problem is easily understood in terms of the geometry of the dual problem, that is, whether or not solutions to the dual problem reside along the edges or faces of the polyhedron. More refined details about *how* sparse x_* needs to be in order to have a reasonable hope of exact recovery require more work, but elementary convex analysis already provides the essential intuition.

For the function $\|x\|_0$ ($p = 0$ in (15)) the equivalence of the primal and dual problems is lost due to the nonconvexity of the objective. The theory of Fenchel duality still yields *weak duality*, but this is of limited use in this instance. The Fenchel dual to (15) is

$$\begin{array}{ll} \underset{y\in\mathbb{R}^m}{\text{maximize}} & b^T y \\ \text{subject to} & (A^* y)_j = 0 \quad j = 1,2,\ldots,n. \end{array} \tag{17}$$

Denoting the *values* of the primal (15) and dual problems (17) by p and d, respectively, these values satisfy the *weak duality inequality* $p \geq d$. The primal problem is a combinatorial optimization problem, and hence NP-hard; the dual problem, however, is a linear program, which is finitely terminating. Relatively elementary variational analysis provides a lower bound on the sparsity of signals x that satisfy the measurements. In this instance, however, the lower bound only reconfirms what is already known. Indeed, if A is full rank, then the only solution to the dual problem is $y = 0$. In other words, the minimal sparsity of the solution to the primal problem is zero, which is obvious. The loss of information in passing from primal to dual formulations of nonconvex problems is a common phenomenon and underscores the importance of convexity.

The Fenchel conjugates of the ℓ_1 norm and the function $\|\cdot\|_0$ are given respectively by

$$\varphi_1^*(y) := \begin{cases} 0 & \|y\|_\infty \le 1 \\ +\infty & \text{else} \end{cases} \qquad (\varphi_1(x) := \|x\|_1) \tag{18}$$

$$\varphi_0^*(y) := \begin{cases} 0 & y = 0 \\ +\infty & \text{else} \end{cases} \qquad (\varphi_0(x) := \|x\|_0) \tag{19}$$

It is not uncommon to consider the function $\|\cdot\|_0$ as the limit of $\left(\sum_j |x_j|^p\right)^{1/p}$ as $p \to 0$. This suggests an alternative approach based on the regularization of the conjugates. For L and $\epsilon > 0$ define

$$\varphi_{\epsilon,L}(y) := \begin{cases} \epsilon\left(\frac{(L+y)\ln(L+y)+(L-y)\ln(L-y)}{2L\ln(2)} - \frac{\ln(L)}{\ln(2)}\right) & (y \in [-L, L]) \\ +\infty & \text{for } |y| > L. \end{cases} \tag{20}$$

This is a scaled and shifted Fermi–Dirac entropy (4). It is also a smooth convex function on the interior of its domain and so elementary calculus can be used to calculate the Fenchel conjugate,

$$\varphi_{\epsilon,L}^*(x) = \frac{\epsilon}{\ln(2)} \ln\left(4^{xL/\epsilon} + 1\right) - xL - \epsilon. \tag{21}$$

For $L > 0$ fixed,

$$\lim_{\epsilon\to 0} \varphi_{\epsilon,L}(y) = \begin{cases} 0 & y \in [-L, L] \\ +\infty & \text{else} \end{cases} \quad \text{and} \quad \lim_{\epsilon\to 0} \varphi_{\epsilon,L}^*(x) = L|x|.$$

For $\epsilon > 0$ fixed

$$\lim_{L\to 0} \varphi_{\epsilon,L}(x) = \begin{cases} 0 & y = 0 \\ +\infty & \text{else} \end{cases} \quad \text{and} \quad \lim_{L\to 0} \varphi_{\epsilon,L}^*(x) := 0.$$

Note that $\|\cdot\|_0$ and $\varphi_{\epsilon,0}^* := 0$ have the same conjugate, but unlike $\|\cdot\|_0$ the biconjugate of $\varphi_{\epsilon,0}^*$ is itself. Also note that $\varphi_{\epsilon,L}$ and $\varphi_{\epsilon,L}^*$ are convex and smooth on the interior of their domains for all $\epsilon, L > 0$. This is in contrast to metrics of the form $\left(\sum_j |x_j|^p\right)$ which are nonconvex for $p < 1$. This suggests solving

$$\begin{array}{ll} \underset{x\in\mathbb{R}^n}{\text{minimize}} & I_{\varphi_{\epsilon,L}^*}(x) \\ \text{subject to} & Ax = b \end{array} \tag{22}$$

as a smooth convex relaxation of the conventional ℓ_p optimization for $0 \leq p \leq 1$. For further details see [28].

Image Denoising and Deconvolution

Consider next problems of the form

$$\underset{x \in X}{\text{minimize}}\ I_\varphi(x) + \frac{1}{2\lambda}\|Ax - y\|^2 \tag{23}$$

where X is a Hilbert space, $I_\varphi : X \to (-\infty, +\infty]$ is a semi-norm on X, and $A : X \to Y$, is a bounded linear operator. This problem is explored in [17] as a general framework that includes total variation minimization [108], wavelet shrinkage [59], and basis pursuit [47]. When A is the identity, problem (23) amounts to a technique for *denoising*; here y is the received, noisy signal, and the solution $\overline{x}$ is an approximation with the desired statistical properties promoted by the objective I_φ. When the linear mapping A is not the identity (for instance, A models convolution against the point spread function of an optical system) problem (23) is a variational formulation of *deconvolution*, that is, recovering the true signal from the image y. The focus here is on total variation minimization.

Total variation was first introduced by Rudin et al. [108] as a regularization technique for denoising images while preserving edges and, more precisely, the statistics of the noisy image. The *total variation* of an image $x \in X = L_1(T)$ – for T and open subset of $\mathbb{R}^2$ – is defined by

$$I_{TV}(x) := \sup\left\{ \int_T x(t)\,\mathrm{div}\,\xi(t)dt \;\middle|\; \xi \in C_c^1(T, \mathbb{R}^2), |\xi(t)| \leq 1\ \forall t \in T \right\}.$$

The integral functional I_{TV} is finite if and only if the distributional derivative Dx of x is a finite Radon measure in T, in which case $I_{TV}(x) = |Dx|(T)$. If, moreover, x has a gradient $\nabla x \in L_1(T, \mathbb{R}^2)$, then $I_{TV}(x) = \int |\nabla x(t)|dt$, or, in the context of the general framework established at the beginning of this chapter, $I_{TV}(x) = I_\varphi(x)$ where $\varphi(x(t)) := |\nabla x(t)|$. The goal of the original *total variation denoising problem* proposed in [108] is then to

$$\begin{array}{ll} \underset{x \in X}{\text{minimize}} & I_{TV}(x) \\ \text{subject to} & \int_T Ax = \int_T x_0 \quad \text{and} \quad \int_T |Ax - x_0|^2 = \sigma^2. \end{array} \tag{24}$$

The first constraint corresponds to the assumption that the noise has zero mean and the second assumption requires the denoised image to have a predetermined standard deviation σ. Under reasonable assumptions [44], this problem is equivalent to the convex optimization problem

$$\begin{array}{ll} \underset{x\in X}{\text{minimize}} & I_{TV}(x) \\ \text{subject to} & \|Ax - x_0\|^2 \le \sigma^2. \end{array} \tag{25}$$

Several authors have exploited duality in total variation minimization for efficient algorithms to solve the above problem [43, 46, 57, 70]. One can "compute" the Fenchel conjugate of I_{TV} indirectly by using the already mentioned property that the *biconjugate* of a proper, convex lsc function is the function itself: $f^{**}(x) = f(x)$ if (and only if) f is proper, convex, and lsc at x. Rewriting I_{TV} as the Fenchel conjugate of some function yields

$$I_{TV}(x) = \sup_v \langle x,\ v\rangle - \iota_K(v),$$

where

$$K := \overline{\{\operatorname{div}\xi \mid \xi \in C_c^1(T,\mathbb{R}^2) \quad \text{and} \quad |\xi(t)| \le 1\ \forall t \in T\}}.$$

From this, it is then clear that the Fenchel conjugate of I_{TV} is the indicator function of the convex set K, ι_K.

In [43], duality is used to develop an algorithm, with proof of convergence, for the problem

$$\underset{x\in X}{\text{minimize}}\ I_{TV}(x) + \frac{1}{2\lambda}\|x - x_0\|^2 \tag{26}$$

with X a Hilbert space. First-order optimality conditions for this unconstrained problem are

$$0 \in x - x_0 + \lambda \partial I_{TV}(x), \tag{27}$$

where $\partial I_{TV}(x)$ is the *subdifferential* of I_{TV} at x defined by

$$v \in \partial I_{TV}(x) \iff I_{TV}(y) \ge I_{TV}(x) + \langle v,\ y - x\rangle \quad \forall y.$$

The optimality condition (27) is equivalent to [24, Proposition 4.4.5]

$$x \in \partial I_{TV}^*\left((x_0 - x)/\lambda\right) \tag{28}$$

or, since $I_{TV}^* = \iota_K$,

$$\frac{x_0}{\lambda} \in \left(I + \frac{1}{\lambda}\partial \iota_K\right)(z)$$

where $z = (x_0 - x)/\lambda$. (For the finite dimensional statement, see [71, Proposition I.6.1.2].) Since K is convex, standard facts from convex analysis determine that $\partial \iota_K(z)$ is the *normal cone mapping* to K at z, denoted $N_K(z)$ and defined by

$$N_K(z) := \begin{cases} \{v \in X \mid \langle v,\ x - z\rangle \le 0 \quad \text{for all } x \in K\} & z \in K \\ \emptyset & z \notin K. \end{cases}$$

Note that this is a set-valued mapping. The *resolvent* $\left(I + \frac{1}{\lambda}\partial\iota_K\right)^{-1}$ evaluated at x_0/λ is the orthogonal *projection* of x_0/λ onto K. That is, the solution to (26) is

$$x_* = x_0 - P_K(x_0/\lambda) = x_0 - P_{\lambda K}(x_0).$$

The inclusions disappear from the formulation due to convexity of K: the resolvent of the normal cone mapping of a convex set is single valued. The numerical algorithm for solving (26) then amounts to an algorithm for computing the projection $P_{\lambda K}$. The tools from convex analysis used in this derivation are the subject of this chapter.

Inverse Scattering

An important problem in applications involving scattering is the determination of the shape and location of scatterers from measurements of the scattered field at a distance. Modern techniques for solving this problem use *indicator functions* to detect the inconsistency or insolubility of an Fredholm integral equation of the first kind, parameterized by points in space. The shape and location of the object is determined by those points where the auxiliary problem is solvable. Equivalently, the technique determines the shape and location of the scatterer by determining whether a sampling function, parameterized by points in space, is in the range of a compact linear operator constructed from the scattering data.

These methods have enjoyed great practical success since their introduction in the latter half of the 1990s. Recently Kirsch and Grinberg [74] established a variational interpretation of these ideas. They observe that the range of a linear operator $G : X \to Y$ (X and Y are reflexive Banach spaces) can be characterized by the infimum of the mapping

$$h(\psi) : Y^* \to \mathbb{R} \cup \{-\infty, +\infty\} := |\langle \psi, F\psi\rangle|,$$

where $F := GSG^*$ for $S : X^* \to X$, a coercive bounded linear operator. Specifically, they establish the following.

Theorem 2 ([74, Theorem 1.16]). *Let X, Y be reflexive Banach spaces with duals X^* and Y^*. Let $F : Y^* \to Y$ and $G : X \to Y$ be bounded linear operators with $F = GSG^*$ for $S : X^* \to X$ a bounded linear operator satisfying the coercivity condition*

$$|\langle \varphi,\ S\varphi\rangle| \ge c\|\varphi\|^2_{X^*} \quad \text{for some } c > 0 \text{ and all } \varphi \in \text{range}(G^*) \subset X^*.$$

Then for any $\phi \in Y \backslash \{0\}$ $\phi \in \text{range}(G)$ *if and only if*

$$\inf\{h(\psi) \mid \psi \in Y^*, \langle \phi,\ \psi \rangle = 1\} > 0.$$

It is shown below that the infimal characterization above is equivalent to the computation of the effective domain of the Fenchel conjugate of h,

$$h^*(\phi) := \sup_{\psi \in Y^*} \{\langle \phi,\ \psi \rangle - h(\psi)\}. \tag{29}$$

In the case of scattering, the operator F above is an integral operator whose kernel is made up of the "measured" field on a surface surrounding the scatterer. When the measurement surface is a sphere at infinity, the corresponding operator is known as the *far field operator*. The factor G maps the boundary condition of the governing PDE (the Helmholtz equation) to the *far field pattern*, that is, the kernel of the far field operator. Given the right choice of spaces, the mapping G is compact, one-to-one, and dense. There are two keys to using the above facts for determining the shape and location of scatterers: first, the construction of the test function ϕ and, second, the connection of the range of G to that of some operator easily computed from the far field operator F. The secret behind the success of these methods in inverse scattering is, first, that the construction of ϕ is trivial and, second, that there is (usually) a simpler object to work with than the infimum in Theorem 2 that depends only on the far field operator (usually the only thing that is known). Indeed, the test functions ϕ are simply far field patterns due to point sources: $\phi_z := e^{-ik\hat{x}\cdot z}$, where $\hat{x}$ is a point on the unit sphere (the direction of the incident field), k is a nonnegative integer (the wave number of the incident field), and z is some point in space.

The crucial observation of Kirsch is that ϕ_z is in the range of G *if and only if* z is a point *inside* the scatterer. If one does not know where the scatter is, let alone its shape, then one does not know G, however, the Fenchel conjugate depends not on G but on the operator F which is constructed from measured data. In general, the Fenchel conjugate, and hence the Kirsch–Grinberg infimal characterization, is difficult to compute, but depending on the physical setting, there is a functional U of F under which the ranges of $U(F)$ and G coincide. In the case where F is a normal operator, $U(F) = (F^*F)^{1/4}$; for non-normal F, the functional U depends more delicately on the physical problem at hand and is only known in a handful of cases. So the algorithm for determining the shape and location of a scatterer amounts to determining those points z, where $e^{-ik\hat{x}\cdot z}$ is in the range of $U(F)$ and where U and F are known and easily computed.

Fredholm Integral Equations

In the scattering application of the previous section, the prevailing numerical technique is not to calculate the Fenchel conjugate of $h(\psi)$ but rather to explore

the range of some functional of F. Ultimately, the computation involves solving a Fredholm integral equation of the first kind, returning to the more general setting with which this chapter began. Let

$$(Ax)(s) = \int_T a(s,t)\mu(dt) = b(s)$$

for reasonable kernels and operators. If A is compact, for instance, as in most deconvolution problems of interest, the problem is *ill posed* in the sense of Hadamard. Some sort of *regularization* technique is therefore required for numerical solutions [60,66,67,76,113]. Regularization is explored in relation to the constraint qualifications (9) or (10).

Formulating the integral equation as an entropy minimization problem yields

$$\begin{array}{ll} \underset{x\in X}{\text{minimize}} & I_\varphi(x) \\ \text{subject to} & Ax = b. \end{array} \tag{30}$$

Following [22, Example 2.2], let T and S be the interval $[0,1]$ with Lebesgue measures μ and ν, and let $a(s,t)$ be a continuous kernel of the Fredholm operator A mapping $X := C([0,1])$ to $Y := C([0,1])$, both equipped with the supremum norm. The adjoint operator is given by $A^*y = \left\{\int_S a(s,t)\lambda(ds)\right\}\mu(dt)$, where the dual spaces are the spaces of Borel measures, $X^* = M([0,1])$ and $Y^* = M([0,1])$. Every element of the range is therefore μ-absolutely continuous and A^* can be viewed as having its range in $L_1([0,1],\mu)$. It follows from [105] that the Fenchel dual of (30) for the operator A is therefore

$$\max_{y^*\in Y^*} \langle b,\ y^*\rangle - I_{\varphi^*}(A^*y^*). \tag{31}$$

Note that the dual problem, unlike the primal, is *unconstrained*. Suppose that A is injective and that $b \in \text{range}(A)$. Assume also that φ^* is everywhere finite and differentiable. Assuming the solution $\overline{y}$ to the dual is attained, the naive application of calculus provides that

$$b = A\left(\frac{\partial\varphi^*}{\partial r}(A^*\overline{y})\right) \quad \text{and} \quad x_\varphi = \left(\frac{\partial\varphi^*}{\partial r}(A^*\overline{y})\right). \tag{32}$$

Similar to the counterexample explored in section "Linear Inverse Problems with Convex Constraints", it is quite likely that $A\left(\frac{\partial\varphi^*}{\partial r}(\text{range}(A^*))\right)$ is smaller than the range of A, hence it is possible to have $b \in \text{range}(A)$ but not in $A\left(\frac{\partial\varphi^*}{\partial r}(\text{range}(A^*))\right)$. Thus the assumption that the solution to the dual problem is attained cannot hold and the primal–dual relationship is broken.

For a specific example, following [22, Example 2.2], consider the Laplace transform restricted to $[0,1]$: $a(s,t) := e^{-st}$ $(s \in [0,1])$, and let φ be either the Boltzmann–Shannon entropy, Fermi–Dirac entropy, or an L_p norm with $p \in (1,2)$,

(3)–(5), respectively. Take $b(s) := \int_{[0,1]} e^{-st}\overline{x}(t)dt$ for $\overline{x} := \alpha\left|t - \frac{1}{2}\right| + \beta$, a solution to (30). It can be shown that the restricted Laplace operator defines an injective linear operator from $C([0,1])$ to $C([0,1])$. However, x_φ given by (32) is continuously differentiable and thus cannot match the known solution $\overline{x}$ which is not differentiable. Indeed, in the case of the Boltzmann–Shannon entropy, the conjugate function and $A^*\overline{y}$ are entire hence the ostensible solution x_φ must be *infinitely* differentiable on $[0,1]$. One could guarantee that the solution to the primal problem (30) is attained by replacing $C([0,1])$ with $L_p([0,1])$, but this does not resolve the problem of attainment in the dual problem.

To recapture the correspondence between primal and dual problems it is necessary to regularize or, alternatively, relax the problem, or to require the constraint qualification $b \in \operatorname{core}(A \operatorname{dom} \varphi)$. Such conditions usually require A to be surjective, or at least to have closed range.

2 Background

As this is meant to be a survey of some of the more useful milestones in convex analysis, the focus is more on the connections between ideas than their proofs. The reader will find the proofs in a variety of sources. The presentation is by default in a normed space X with dual X^*, though if statements become too technical, the Euclidean space variants will suffice. E denotes a finite-dimensional real vector space $\mathbb{R}^n$ for some $n \in \mathbb{N}$ endowed with the usual norm. Typically, X will be reserved for a real infinite-dimensional Banach space. A common convention in convex analysis is to include one or both of $-\infty$ and $+\infty$ in the range of functions (typically only $+\infty$). This is denoted by the (semi-) closed interval $(-\infty, +\infty]$ or $[-\infty, +\infty]$.

A set $C \subset X$ is said to be convex if it contains all line segments between any two points in C: $\lambda x + (1-\lambda)y \in C$ for all $\lambda \in [0,1]$ and $x, y \in C$. Much of the theory of convexity is centered on the analysis of convex sets, however, sets and functions are treated interchangeably through the use of level sets, epigraphs, and indicator functions. The *lower-level sets* of a function $f : X \to [-\infty, +\infty]$ are denoted $\operatorname{lev}_{\leq\alpha} f$ and defined by $\operatorname{lev}_\alpha f := \{x \in X \mid f(x) \leq \alpha\}$ where $\alpha \in \mathbb{R}$. The *epigraph* of a function $f : X \to [-\infty, +\infty]$ is defined by

$$\operatorname{epi} f := \{(x,t) \in E \times \mathbb{R} \mid f(x) \leq t\}.$$

This leads to the very natural definition of a *convex function* as one whose epigraph is a convex set. More directly, a convex function is defined as a mapping $f : X \to [-\infty, +\infty]$ with convex domain and

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y) \quad \text{for any } x, y \in \operatorname{dom} f \text{ and } \lambda \in [0,1].$$

A proper convex function $f : X \to [-\infty, +\infty]$ is *strictly* convex if the above inequality is strict for all distinct x and y in the domain of f and all $0 < \lambda < 1$.

A function is said to be *closed* if its epigraph is closed; whereas a *lower semi-continuous* (lsc) function f satisfies $\liminf_{x\to\overline{x}} f(x) \geq f(\overline{x})$ for all $\overline{x} \in X$. These properties are in fact equivalent:

Proposition 1. *The following properties of a function $f : X \to [-\infty, +\infty]$ are equivalent:*

(i) *f is lsc.*
(ii) *epi f is closed in $X \times \mathbb{R}$.*
(iii) *The level sets $\operatorname{lev}_{\leq\alpha} f$ are closed on X for each $\alpha \in \mathbb{R}$.*

Guide. For Euclidean spaces, this is shown in [107, Theorem 1.6]. In the Banach space setting this is [24, Proposition 4.1.1]. This is left as an exercise for the Hilbert space setting in [50, Exercise 2.1]. ■

A principal focus is on *proper* functions, that is, $f : E \to [-\infty, +\infty]$ with nonempty domain. The indicator function is often used to pass from sets to functions

$$\iota_C(x) := \begin{cases} 0 & x \in C \\ +\infty & \text{else.} \end{cases}$$

For $C \subset X$ convex, $f : C \to [-\infty, +\infty]$ will be referred to as a convex function if the extended function

$$\overline{f}(x) := \begin{cases} f(x) & x \in C \\ +\infty & \text{else} \end{cases}$$

is convex.

Lipschitzian Properties

Convex functions have the remarkable, yet elementary, property that local boundedness and local Lipschitz properties are *equivalent* without any additional assumptions on the function. In the following statement of this fact, the unit ball is denoted by $B_X := \{x \in X \mid \|x\| \leq 1\}$.

Lemma 1. *Let $f : X \to (-\infty, +\infty]$ be a convex function and suppose that $C \subset X$ is a bounded convex set. If f is bounded on $C + \delta B_X$ for some $\delta > 0$, then f is Lipschitz on C.*

Guide. See [24, Lemma 4.1.3]. ■

With this fact, one can easily establish the following.

Proposition 2 (Convexity and Continuity in Normed Spaces). *Let $f : X \to (-\infty, +\infty]$ be proper and convex, and let $x \in \operatorname{dom} f$. The following are equivalent:*

(i) *f is Lipschitz on some neighborhood of x.*
(ii) *f is continuous at x.*
(iii) *f is bounded on a neighborhood of x.*
(iv) *f is bounded above on a neighborhood of x.*

Guide. See [24, Proposition 4.1.4] or [31, Sect. 4.1.2]. ■

In finite dimensions, convexity and continuity are much more tightly connected.

Proposition 3 (Convexity and Continuity in Euclidean Spaces). *Let $f : E \to (-\infty, +\infty]$ be convex. Then f is locally Lipschitz, and hence continuous, on the interior of its domain.*

Guide. See [24, Theorem 2.1.12] or [72, Theorem 3.1.2] ■

Unlike finite dimensions, in infinite dimensions a convex function need not be continuous. A Hamel basis, for instance, an algebraic basis for the vector space can be used to define discontinuous linear functionals [24, Exercise 4.1.21]. For lsc convex functions, however, the correspondence follows through. The following statement uses the notion of the *core* of a set given by Definition 1.

Example 1 (A Discontinuous Linear Functional). Let c_{00} denote the normed subspace of all finitely supported sequences in c_0, the vector space of sequences in X converging to 0; obviously c_{00} is open. Define $\Lambda : c_{00} \to \mathbb{R}$ by $\Lambda(x) = \sum x_j$ where $x = (x_j) \in c_{00}$. This is clearly a linear functional and discontinuous at 0. Now extend Λ to a functional $\hat{\Lambda}$ on the Banach space c_0 by taking a basis for c_0 considered as a vector space over c_{00}. In particular, $C := \hat{\Lambda}^{-1}([-1, 1])$ is a convex set with empty interior for which 0 is a core point. Moreover, $\overline{C} = c_0$ and $\hat{\Lambda}$ is certainly discontinuous. ■

Proposition 4 (Convexity and Continuity in Banach Spaces). *Suppose X is a Banach space and $f : X \to (-\infty, +\infty]$ is lsc, proper, and convex. Then the following are equivalent:*

(i) *f is continuous at x.*
(ii) *$x \in \operatorname{int} \operatorname{dom} f$.*
(iii) *$x \in \operatorname{core} \operatorname{dom} f$.*

Guide. This is [24, Theorem 4.1.5]. See also [31, Theorem 4.1.3]. ■

The above result is helpful since it is often easier to verify that a point is in the core of the domain of a convex function than in the interior.

Subdifferentials

The analog to the linear function in classical analysis is the *sublinear function* in convex analysis. A function $f : X \to [-\infty, +\infty]$ is said to be *sublinear* if

$$f(\lambda x + \gamma y) \leq \lambda f(x) + \gamma f(y) \quad \text{for all } x, y \in X \text{ and } \lambda, \gamma \geq 0.$$

By convention, $0 \cdot (+\infty) = 0$. Sometimes sublinearity is defined as a function f that is *positively homogeneous (of degree 1)* – i.e., $0 \in \operatorname{dom} f$ and $f(\lambda x) = \lambda f(x)$ for all x and all $\lambda > 0$ – and is *subadditive*

$$f(x + y) \leq f(x) + f(y) \quad \text{for all } x \text{ and } y.$$

Example 2 (Norms). A *norm* on a vector space is a sublinear function. Recall that a nonnegative function $\|\cdot\|$ on a vector space X is a norm if

(i) $\|x\| \geq 0$ for each $x \in X$.
(ii) $\|x\| = 0$ if and only if $x = 0$.
(iii) $\|\lambda x\| = |\lambda| \|x\|$ for every $x \in X$ and scalar λ.
(iv) $\|x + y\| \leq \|x\| + \|y\|$ for every $x, y \in X$.

A *normed space* is a vector space endowed with such a norm and is called a *Banach space* if it is *complete* which is to say that all Cauchy sequences converge. ■

Another important sublinear function is the *directional derivative* of the function f at x in the direction d defined by

$$f'(x; d) := \lim_{t \searrow 0} \frac{f(x + td) - f(x)}{t}$$

whenever this limit exists.

Proposition 5 (Sublinearity of the Directional Derivative). *Let X be a Banach space and let $f : X \to (-\infty, +\infty]$ be a convex function. Suppose that $\overline{x} \in \operatorname{core}(\operatorname{dom} f)$. Then the directional derivative $f'(\overline{x}; \cdot)$ is everywhere finite and sublinear.*

Guide. See [31, Proposition 4.2.4]. For the finite dimensional analog, see [72, Proposition D.1.1.2] or [24, Proposition 2.1.17]. ■

Another important instance of sublinear functions are *support functions* of convex sets, which, in turn permit local first-order approximations to convex functions. A *support function* of a nonempty subset S of the dual space X^*, usually denoted σ_S, is defined by $\sigma_S(x) := \sup\{\langle s, x\rangle \mid s \in S\}$. The support function is convex, proper (not everywhere infinite), and $0 \in \operatorname{dom} \sigma_S$.

Example 3 (Support Functions and Fenchel Conjugation). From the definition of the support function it follows immediately that, for a closed convex set C,

$$\iota_C^* = \sigma_C \quad \text{and} \quad \iota_C^{**} = \iota_C. \qquad \blacksquare$$

A powerful observation is that any closed sublinear function can be viewed as a support function. This can be seen by the representation of closed convex functions via affine minorants. This is the content of the *Hahn–Banach* theorem, which is stated in infinite dimensions as this setting will be needed below.

Theorem 3 (Hahn–Banach: Analytic Form). *Let X be a normed space and $\sigma : X \to \mathbb{R}$ be a continuous sublinear function with* dom $\sigma = X$. *Suppose that L is a linear subspace of X and that the linear function $h : L \to \mathbb{R}$ is dominated by σ on L, that is $\sigma \geq h$ on L. Then there is a linear function minorizing σ on X, that is, there exists a $x^* \in X^*$ dominated by σ such that $h(x) = \langle x^*, x\rangle \leq \sigma(x)$ for all $x \in L$.*

Guide. The proof can be carried out in finite dimensions with elementary tools, constructing x^* from h sequentially by one-dimensional extensions from L. See [72, Theorem C.3.1.1] and [24, Proposition 2.1.18]. The technique can be extended to Banach spaces using Zorn's lemma and a verification that the linear functionals so constructed are continuous (guaranteed by the domination property) [24, Theorem 4.1.7]. See also [110, Theorem 1.11]. $\blacksquare$

An important point in the Hahn–Banach extension theorem is the *existence* of a minorizing linear function, and hence the existence of the *set* of linear minorants. In fact, σ is the supremum of the linear functions minorizing it. In other words, σ is the support function of the nonempty set

$$S_\sigma := \{s \in X^* \mid \langle s, x\rangle \leq \sigma(x) \quad \text{for all } x \in X\}.$$

A number of facts follow from Theorem 3, in particular the nonemptiness of the subdifferential, a sandwich theorem and, thence, Fenchel Duality (respectively, Theorems 5, 7, and 12). It turns out that the converse also holds, and thus these facts are actually *equivalent* to nonemptiness of the subdifferential. This is the so-called *Hahn–Banach/Fenchel duality* circle.

As stated in Proposition 5, the directional derivative is everywhere finite and sublinear for a convex function f at points in the core of its domain. In light of the Hahn–Banach theorem, then $f'(\overline{x}, \cdot)$ can be expressed for all $d \in X$ in terms of its minorizing function:

$$f'(\overline{x}, d) = \sigma_S(d) = \max_{v \in S}\{\langle v, d\rangle\}.$$

The set S for which $f'(\overline{x}, d)$ is the support function has a special name: the *subdifferential* of f at $\overline{x}$. It is tempting to *define* the subdifferential this way, however,

there is a more elemental definition that does not rely on directional derivatives or support functions, or indeed even the convexity of f. The correspondence between directional derivatives of convex functions and the subdifferential below is a consequence of the Hahn–Banach theorem.

Definition 2 (Subdifferential). For a function $f : X \to (-\infty, +\infty]$ and a point $\overline{x} \in \operatorname{dom} f$, the *subdifferential* of f at $\overline{x}$, denoted $\partial f(\overline{x})$ is defined by

$$\partial f(\overline{x}) := \{v \in X^* \mid v(x) - v(\overline{x}) \leq f(x) - f(\overline{x}) \text{ for all } x \in X\}.$$

When $\overline{x} \notin \operatorname{dom} f$, define $\partial f(\overline{x}) = \emptyset$.

In Euclidean space the subdifferential is just

$$\partial f(\overline{x}) = \{v \in E \mid \langle v, x\rangle - \langle v, \overline{x}\rangle \leq f(x) - f(\overline{x}) \text{ for all } x \in E\}.$$

An element of $\partial f(x)$ is called a *subgradient* of f at x. See [31,93,107] for more in-depth discussion of the regular, or limiting subdifferential defined here, in addition to other useful varieties. This is a generalization of the classical gradient. Just as the gradient need not exist, the subdifferential of a lsc convex function may be empty at some points in its domain. Take, for example, $f(x) = -\sqrt{1-x^2}$ for $-1 \leq x \leq 1$. Then $\partial f(x) = \emptyset$ for $x = \pm 1$.

Example 4 (Common Subdifferentials).

(i) Gradients. A function $f : X \to \mathbb{R}$ is said to be *strictly differentiable* at $\overline{x}$ if

$$\lim_{x \to \overline{x}, u \to \overline{x}} \frac{f(x) - f(u) - \nabla f(\overline{x})(x-u)}{\|x-u\|} = 0.$$

This is a stronger differentiability property than Fréchet differentiability since it requires uniformity in *pairs* of points converging to $\overline{x}$. Luckily for convex functions the two notions agree. If f is convex and strictly differentiable at $\overline{x}$, then the subdifferential is exactly the gradient. (This follows from the equivalence of the subdifferential in Definition 2 and the basic limiting subdifferential defined in [93, Definition 1.77] for convex functions and [93, Corollary 1.82].) In finite dimensions, at a point $\overline{x} \in \operatorname{dom} f$ for f convex, Fréchet and Gâteaux differentiability coincide, and the subdifferential is a singleton [24, Theorem 2.2.1]. In infinite dimensions, a convex function f that is continuous at $\overline{x}$ is Gâteaux differentiable at $\overline{x}$ if and only if the $\partial f(\overline{x})$ is a singleton [24, Corollary 4.2.11].

(ii) The subdifferential of the indicator function.

$$\partial \iota_C(\overline{x}) = N_C(\overline{x}),$$

where $C \subset X$ is closed and convex, X is a Banach, and $N_C(\overline{x}) \subset X^*$ is the *normal cone mapping* to C at $\overline{x}$ defined by

$$N_C(\overline{x}) := \begin{cases} \{v \in X^* \mid \langle v,\ x - \overline{x}\rangle \leq 0 \quad \text{for all } x \in C\} & \overline{x} \in C \\ \emptyset & \overline{x} \notin C. \end{cases} \tag{33}$$

See (41) for alternative definitions and further discussion of this important mapping.

(iii) Absolute value. For $x \in \mathbb{R}$,

$$\partial|\cdot|(x) = \begin{cases} -1 & x < 0 \\ [-1, 1] & x = 0 \\ 1 & x > 0. \end{cases}$$

■

The following elementary observation suggests the fundamental significance of subdifferential in optimization.

Theorem 4 (Subdifferential at Optimality: Fermat's Rule). *Let X be a normed space, and let $f : X \to (-\infty, +\infty]$ be proper and convex. Then f has a (global) minimum at $\overline{x}$ if and only if $0 \in \partial f(\overline{x})$.*

Guide. The first implication of the global result follows from a more general local result [93, Proposition 1.114] by convexity; the converse statement follows from the definition of the subdifferential and convexity. ■

Returning now to the correspondence between the subdifferential and the directional derivative of a convex function $f'(x; d)$ has the following fundamental result.

Theorem 5 (Max Formula – Existence of ∂f). *Let X be a normed space, $d \in X$ and let $f : X \to (-\infty, +\infty]$ be convex. Suppose that $\overline{x} \in \text{cont } f$. Then $\partial f(\overline{x}) \neq \emptyset$ and*

$$f'(\overline{x}, d) = \max\{\langle x^*,\ d\rangle \mid x^* \in \partial f(\overline{x})\}.$$

Proof. The tools are in place for a simple proof that synthesizes many of the facts tabulated so far. By Proposition 5 $f'(\overline{x}; \cdot)$ is finite; so, for fixed $d \in \{x \in X \mid \|x\| = 1\}$, let $\alpha = f'(\overline{x}; d) < \infty$. The stronger assumption that $\overline{x} \in \text{cont } f$ and the convexity of $f'(\overline{x}; \cdot)$ yield that the directional derivative is Lipschitz continuous with constant K. Let $S := \{td \mid t \in \mathbb{R}\}$ and define the linear function $\Lambda : S \to \mathbb{R}$ by $\Lambda(td) := t\alpha$ for $t \in \mathbb{R}$. Then $\Lambda(\cdot) \leq f'(\overline{x}; \cdot)$ on S. The Hahn–Banach theorem 3 then guarantees the existence of $\phi \in X^*$ such that

$$\phi = \Lambda \text{ on } S, \quad \phi(\cdot) \leq f'(\overline{x}; \cdot) \text{ on } X.$$

Then $\phi \in \partial f(\overline{x})$ and $\phi(sd) = f'(\overline{x}; sd)$ for all $s \geq 0$. ■

A simple example on $\mathbb{R}$ illustrates the importance of the qualification $\overline{x} \in \text{cont } f$. Let

$$f(x) : \mathbb{R} \to (-\infty, +\infty] := \begin{cases} -\sqrt{x}, & x \geq 0 \\ +\infty & \text{otherwise.} \end{cases}$$

For this example, $\partial f(0) = \emptyset$.

An important application of the Max formula in finite dimensions is the mean value theorem for convex functions.

Theorem 6 (Convex Mean Value Theorem). *Let $f : E \to (-\infty, +\infty]$ be convex and continuous. For $u, v \in E$ there exists a point $z \in E$ interior to the line segment $[u, v]$ with*

$$f(u) - f(v) \leq \langle w,\ u - v \rangle \quad \textit{for all } w \in \partial f(z).$$

Guide. See [93, 107] for extensions of this result and detailed historical background. ■

The next theorem is a key tool in developing a subdifferential calculus. It relies on assumptions that are used frequently enough that it is worthwhile to present them separately.

Assumption 6. *Let X and Y be Banach spaces and let $T : X \to Y$ be a bounded linear mapping. Let $f : X \to (-\infty, +\infty]$ and $g : Y \to (-\infty, +\infty]$ satisfy one of*

$$0 \in \text{core}\,(\text{dom } g - T \text{ dom } f) \textit{ and both } f \textit{ and } g \textit{ are lsc,} \tag{34}$$

or

$$T \text{ dom } f \cap \text{cont } g \neq \emptyset. \tag{35}$$

The later assumption can be used in incomplete normed spaces as well.

Theorem 7 (Sandwich Theorem). *Let X and Y be Banach spaces and let $T : X \to Y$ be a bounded linear mapping. Suppose that $f : X \to (-\infty, +\infty]$ and $g : Y \to (-\infty, +\infty]$ are proper convex functions with $f \geq -g \circ T$ and which satisfy Assumption 6. Then there is an affine function $A : X \to \mathbb{R}$ defined by $Ax := \langle T^* y^*,\ x \rangle + r$ satisfying $f \geq A \geq -g \circ T$. Moreover, for any $\overline{x}$ satisfying $f(\overline{x}) = (-g \circ T)(\overline{x})$, it holds that $-y^* \in \partial g(T\overline{x})$.*

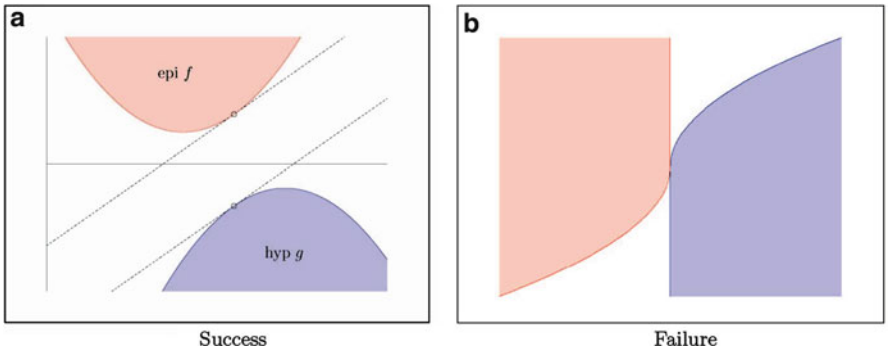


Fig. 1 Hahn–Banach sandwich theorem and its failure. (**a**) Success; (**b**) failure

Guide. By the development to this point, the Max formula [24, Theorem 4.1.18] would apply to prove the result. For a vector space version see [110, Corollary 2.1]. Another route is via Fenchel duality which is explored in the next section. A third approach closely related to the Fenchel duality approach [31, Theorem 4.3.2] is based on a *decoupling* lemma which is also presented in the next section (Lemma 3). ■

Corollary 2 (Basic Separation). *Let $C \subset X$ be a nonempty convex set with nonempty interior in a normed space, and suppose $x_0 \notin \operatorname{int} C$. Then there exists $\phi \in X^* \backslash \{0\}$ such that*

$$\sup_C \phi \leq \phi(x_0) \quad \text{and} \quad \phi(x) < \phi(x_0) \text{ for all } x \in \operatorname{int} C.$$

If $x_0 \notin \overline{C}$ then it can be assumed that $\sup_C \phi < \phi(x_0)$.

Proof. Assume without loss of generality that $0 \in \operatorname{int} C$ and apply the sandwich theorem with $f = \iota_{\{x_0\}}$, T the identity mapping on X and $g(x) = \inf\{r > 0 \mid x \in rC\} - 1$. See [31, Theorem 4.3.8] and [24, Corollary 4.1.15]. ■

The Hahn–Banach theorem 3 can be seen as an easy consequence of the sandwich theorem 7, which completes part of the circle. Figure 1 illustrates these ideas.

In the next section Fenchel duality is added to this cycle. A calculus of subdifferentials and a few fundamental results connecting the subdifferential to classical derivatives and *monotone operators* conclude the present section.

Theorem 8 (Subdifferential Sum Rule). *Let X and Y be Banach spaces, $T : X \to Y$ a bounded linear mapping and let $f : X \to (-\infty, +\infty]$ and $g : Y \to (-\infty, +\infty]$ be convex functions. Then at any point $x \in X$*

$$\partial (f + g \circ T)(x) \supset \partial f(x) + T^*(\partial g(Tx)),$$

with equality if Assumption 6 holds.

Proof sketch. The inclusion is clear. Proving equality permits an elegant proof using the sandwich theorem [24, Theorem 4.1.19], which we sketch here. Take $\phi \in \partial\,(f + g \circ T)\,(\overline{x})$ and assume without loss of generality that

$$x \mapsto f(x) + g(Tx) - \phi(x)$$

attains a minimum of 0 at $\overline{x}$. By Theorem 7 there is an affine function $A := \langle T^* y^*,\ \cdot\rangle + r$ with $-y^* \in \partial g(T\overline{x})$ such that

$$f(x) - \phi(x) \geq Ax \geq -g(Ax).$$

Equality is attained at $x = \overline{x}$. It remains to check that $\phi + T^* y^* \in \partial f(\overline{x})$. ■

The next result is a useful extension to Proposition 2.

Theorem 9 (Convexity and Regularity in Normed Spaces). *Let $f : X \to (-\infty, +\infty]$ be proper and convex, and let $x \in \operatorname{dom} f$. The following are equivalent:*

(i) *f is Lipschitz on some neighborhood of x.*
(ii) *f is continuous at x.*
(iii) *f is bounded on a neighborhood of x.*
(iv) *f is bounded above on a neighborhood of x.*
(v) *∂f maps bounded subsets of X into bounded nonempty subsets of X^*.*

Guide. See [24, Theorem 4.1.25]. ■

The next results relate to Example 4 and provide additional tools for verifying differentiability of convex functions. The notation $\to_{w^*}$ denotes weak* convergence.

Theorem 10 (Šmulian). *Let the convex function f be continuous at $\overline{x}$.*

(i) *The following are equivalent:*
  (a) *f is Fréchet differentiable at $\overline{x}$.*
  (b) *For each sequence $x_n \to \overline{x}$ and $\phi \in \partial f(\overline{x})$, there exist $\overline{n} \in \mathbb{N}$ and $\phi_n \in \partial f(x_n)$ for $n \geq \overline{n}$ such that $\phi_n \to \phi$.*
  (c) *$\phi_n \to \phi$ whenever $\phi_n \in \partial f(x_n)$, $\phi \in \partial f(\overline{x})$.*
(ii) *The following are equivalent:*
  (a) *f is Gâteaux differentiable at $\overline{x}$.*
  (b) *For each sequence $x_n \to \overline{x}$ and $\phi \in \partial f(\overline{x})$, there exist $\overline{n} \in \mathbb{N}$ and $\phi_n \in \partial f(x_n)$ for $n \geq \overline{n}$ such that $\phi_n \to_{w^*} \phi$.*
  (c) *$\phi_n \to_{w^*} \phi$ whenever $\phi_n \in \partial f(x_n)$, $\phi \in \partial f(\overline{x})$.*

A more complete statement of these facts and their provenance can be found in [24, Theorems 4.2.8 and 4.2.9]. In particular, in every infinite dimensional normed space,

there is a continuous convex function which is Gâteaux but not Fréchet differentiable at the origin.

An elementary but powerful observation about the subdifferential viewed as a multi-valued mapping will conclude this section. A multi-valued mapping T from X to X^* is denoted with double arrows, $T : X \rightrightarrows X^*$. Then T is *monotone* if

$$\langle v_2 - v_1, x_2 - x_1 \rangle \geq 0 \quad \text{whenever } v_1 \in T(x_1),\ v_2 \in T(x_2).$$

Proposition 6 (Monotonicity and Convexity). *Let $f : X \to (-\infty, +\infty]$ be proper and convex on a normed space. Then the subdifferential mapping $\partial f : X \rightrightarrows X^*$ is monotone.*

Proof. Add the subdifferential inequalities in the Definition 2 applied to $f(x_1)$ and $f(x_0)$ for $v_1 \in \partial f(x_1)$ and $v_0 \in \partial(f(x_0)$. ■

3 Duality and Convex Analysis

The Fenchel conjugate is to convex analysis what the Fourier transform is to harmonic analysis. To begin, some basic facts about this fundamental tool are collected.

Fenchel Conjugation

The Fenchel conjugate, introduced in [62], of a mapping $f : X \to [-\infty, +\infty]$, as mentioned above is denoted $f^* : X^* \to [-\infty, +\infty]$ and defined by

$$f^*(x^*) = \sup_{x \in X} \{\langle x^*,\ x \rangle - f(x)\}.$$

The conjugate is always convex (as a supremum of affine functions). If the domain of f is nonempty, then f^* never takes the value $-\infty$.

Example 5 (Important Fenchel Conjugates).

(i) Absolute value.

$$f(x) = |x| \quad (x \in \mathbb{R}), \qquad f^*(y) = \begin{cases} 0 & y \in [-1, 1] \\ +\infty & \text{else.} \end{cases}$$

(ii) L_p norms $(p > 1)$.

$$f(x) = \frac{1}{p}\|x\|^p \quad (p > 1), \qquad f^*(y) = \frac{1}{q}\|y\|^q \quad \left(\tfrac{1}{p} + \tfrac{1}{q} = 1\right).$$

In particular, note that the two-norm conjugate is "self-conjugate."

(iii) Indicator functions.

$$f = \iota_C, \quad f^* = \sigma_C,$$

where σ_C is the *support function* of the set C. Note that if C is not closed and convex, then the conjugate of σ_C, that is the *biconjugate* of ι_C, is the *closed convex hull* of C. (See Proposition 8(ii).)

(iv) Boltzmann–Shannon entropy.

$$f(x) = \begin{cases} x \ln x - x & (x > 0) \\ 0 & (x = 0) \end{cases}, \quad f^*(y) = e^y \quad (y \in \mathbb{R}).$$

(v) Fermi–Dirac entropy.

$$f(x) = \begin{cases} x \ln x + (1-x)\ln(1-x) & (x \in (0,1)) \\ 0 & (x = 0, 1) \end{cases},$$
$$f^*(y) = \ln(1 + e^y) \quad (y \in \mathbb{R}).$$

■

Some useful properties of conjugate functions are tabulated below.

Proposition 7 (Fenchel–Young Inequality). *Let X be a normed space and let $f : X \to [-\infty, +\infty]$. Suppose that $x^* \in X^*$ and $x \in \operatorname{dom} f$. Then*

$$f(x) + f^*(x^*) \geq \langle x^*, x \rangle . \tag{36}$$

Equality holds if and only if $x^* \in \partial f(x)$.

Proof sketch. The proof follows by an elementary application of the definitions of the Fenchel conjugate and the subdifferential. See [103] for the finite dimensional case. The same proof works in the normed space setting. ■

The conjugate, as the supremum of affine functions, is convex. In the following, the closure of a function f is denoted by $\overline{f}$, and $\overline{\operatorname{conv}} f$ is the function whose epigraph is the closed convex hull of the epigraph of f.

Proposition 8. *Let X be a normed space and let $f : X \to [-\infty, +\infty]$.*

(i) *If $f \geq g$ then $g^* \geq f^*$.*

(ii) $f^* = \left(\overline{f}\right)^* = (\overline{\operatorname{conv}} f)^*$.

Proof. The definition of the conjugate immediately implies (i). This immediately yields $f^* \leq \left(\overline{f}\right)^* \leq (\overline{\operatorname{conv}} f)^*$. To show (ii) it remains to show that $f^* \geq$

$(\overline{\text{conv}}\, f)^*$. Choose any $\phi \in X^*$. If $f^*(\phi) = +\infty$ the conclusion is clear, so assume $f^*(\phi) = \alpha$ for some $\alpha \in \mathbb{R}$. Then $\phi(x) - f(x) \leq \alpha$ for all $x \in X$. Define $g := \phi - f$. Then $g \leq \overline{\text{conv}}\, f$ and, by (i) $(\overline{\text{conv}}\, f)^* \leq g^*$. But $g^* = \alpha$, so $(\overline{\text{conv}}\, f)^* \leq \alpha = f^*(\phi)$. ■

Application of Fenchel conjugation twice, or *biconjugation* denoted by f^{**}, is a function on X^{**}. In certain instances, biconjugation is the identity – in this way, the Fenchel conjugate resembles the Fourier transform. Indeed, Fenchel conjugation plays a role in the convex analysis similar to the Fourier transform in harmonic analysis and has a contemporaneous provenance dating back to Legendre.

Proposition 9 (Biconjugation). *Let $f : X \to (-\infty, +\infty]$, $x \in X$ and $x^* \in X^*$.*

(i) $f^{**}|_X \leq f$.
(ii) *If f is convex and proper, then $f^{**}(x) = f(x)$ at x if and only if f is lsc at x. In particular, f is lsc if and only if $f_X^{**} = f$.*
(iii) $f^{**}|_X = \overline{\text{conv}}\, f$ *if* $\overline{\text{conv}}\, f$ *is proper.*

Guide. (i) follows from Fenchel–Young, Proposition 7, and the definition of the conjugate. (ii) follows from (i) and an epi-separation property [24, Proposition 4.4.2]. (iii) follows from (ii) of this proposition and 8(ii). ■

The next results highlight the relationship between the Fenchel conjugate and the subdifferential already used in (28).

Proposition 10. *Let $f : X \to (-\infty, +\infty]$ be a function and $\overline{x} \in \text{dom}\, f$. If $\phi \in \partial f(\overline{x})$ then $\overline{x} \in \partial f^*(\phi)$. If, additionally, f is convex and lsc at $\overline{x}$, then the converse holds, namely $\overline{x} \in \partial f^*(\phi)$ implies $\phi \in \partial f(\overline{x})$.*

Guide. See [72, Corollary 1.4.4] for the finite dimensional version of this fact that, with some modification, can be extended to normed spaces. ■

Infimal convolutions conclude this subsection. Among them many applications are smoothing and approximation – just as is the case for integral convolutions.

Definition 3 (Infimal Convolution). Let f and g be proper extended real-valued functions on a normed space X. The *infimal convolution* of f and g is defined by

$$(f \square g)(x) := \inf_{y \in X} f(y) + g(x - y).$$

The infimal convolution of f and g is the largest extended real-valued function whose epigraph contains the sum of epigraphs of f and g; consequently, it is a convex function when f and g are convex.

The next lemma follows directly from the definitions and careful application of the properties of suprema and infima.

Lemma 2. *Let X be a normed space and let f and g be proper functions on X, then $(f \square g)^* = f^* + g^*$.*

An important example of infimal convolution is *Yosida approximation*.

Theorem 11 (Yosida Approximation). *Let $f : X \to \mathbb{R}$ be convex and bounded on bounded sets. Then both $f \square n\|\cdot\|^2$ and $f \square n\|\cdot\|$ converge uniformly to f on bounded sets.*

Guide. This follows from the above lemma and basic approximation facts. ■

In the inverse problems literature $\left(f \square n\|\cdot\|^2\right)(0)$ is often referred to as *Tikhonov regularization*; elsewhere, $f \square n\|\cdot\|^2$ is referred to as *Moreau–Yosida regularization* because $f \square \frac{1}{2}\|\cdot\|^2$, the *Moreau envelope*, was studied in depth by Moreau [94, 95]. The argmin mapping corresponding to the Moreau envelope – that is the mapping of $x \in X$ to the point $\overline{y} \in X$ at which the value of $f \square \frac{1}{2}\|\cdot\|^2$ is attained – is called the *proximal mapping* [94, 95, 107]

$$\operatorname{prox}_{\lambda, f}(x) := \operatorname{argmin}_{y \in X} f(y) + \frac{1}{2\lambda}\|x - y\|^2. \tag{37}$$

When f is the indicator function of a closed convex set C, the proximal mapping is just the *metric projection* onto C, denoted by $P_C(x)$: $\operatorname{prox}_{\lambda, \iota_C}(x) = P_C(x)$.

Fenchel Duality

Fenchel duality can be proved by Theorem 5 and the sandwich theorem 7 [24, Theorem 4.4.18]. As the syllogism to this point deduces Fenchel duality as a consequence of the Hahn–Banach theorem. In order to close the Fenchel duality/Hahn–Banach circle of ideas, however, following [31] the main duality result of this section follows from the Fenchel–Young inequality and the next important lemma.

Lemma 3 (Decoupling). *Let X and Y be Banach spaces and let $T : X \to Y$ be a bounded linear mapping. Suppose that $f : X \to (-\infty, +\infty]$ and $g : Y \to (-\infty, +\infty]$ are proper convex functions which satisfy Assumption 6. Then there is a $y^* \in Y^*$ such that for any $x \in X$ and $y \in Y$,*

$$p \leq \left(f(x) - \langle y^*,\ Tx\rangle\right) + \left(g(y) + \langle y^*,\ y\rangle\right),$$

where $p := \inf_X\{f(x) + g(Tx)\}$.

Guide. Define the perturbed function $h : Y \to [-\infty, +\infty]$ by

$$h(u) := \inf_{x \in X} \{f(x) + g(Tx + u)\},$$

which has the property that h is convex, dom $h =$ dom $g - T$ dom f and (the most technical part of the proof) $0 \in \text{int}\,(\text{dom}\, h)$. This can be proved by assuming the first of the constraint qualifications (34). The second condition (35) implies (34). Then Theorem 5 yields $\partial h(0) \neq \emptyset$, which guarantees the attainment of a minimum of the perturbed function. The decoupling is achieved through a particular choice of the perturbation u. See [31, Lemma 4.3.1]. ■

One can now provide an elegant proof of Theorem 1, which is restated here for convenience.

Theorem 12 (Fenchel Duality). *Let X and Y be normed spaces, consider the functions $f : X \to (-\infty, +\infty]$ and $g : Y \to (-\infty, +\infty]$ and let $T : X \to Y$ be a bounded linear map. Define the primal and dual values $p, d \in [-\infty, +\infty]$ by the Fenchel problems*

$$p = \inf_{x \in X} \{f(x) + g(Tx)\} \tag{38}$$

$$d = \sup_{y^* \in Y^*} \{-f^*(T^* y^*) - g^*(-y^*)\}. \tag{39}$$

These values satisfy the weak duality inequality $p \geq d$.

If X, Y are Banach, f, g are convex and satisfy Assumption 6 then $p = d$, and the supremum to the dual problem is attained if finite.

Proof. Weak duality follows directly from the Fenchel–Young inequality.

For equality assume that $p \neq -\infty$ (this case is clear). Then Assumption 6 guarantees that $p < +\infty$, and by the decoupling lemma (Lemma 3), there is a $\phi \in Y^*$ such that for all $x \in X$ and $y \in Y$

$$p \leq (f(x) - \langle \phi,\ Tx \rangle) + (g(y) - \langle -\phi,\ y \rangle)\,.$$

Taking the infimum over all x and then over all y yields

$$p \leq -f^*(T^*, \phi) - g^*(-\phi) \leq d \leq p.$$

Hence, ϕ attains the supremum in (39), and $p = d$. ■

Fenchel duality for *linear constraints*, Corollary 1, follows immediately by taking $g := \iota_{\{b\}}$.

Applications

Calculus Fenchel duality is, in some sense, the dual space representation of the sandwich theorem. It is a straightforward exercise to derive Fenchel duality from Theorem 7. Conversely, the existence of a point of attainment in Theorem 12 yields an explicit construction of the linear mapping in Theorem 7: $A := \langle T^*\phi,\ \cdot\rangle + r$, where ϕ is the point of attainment in (39) and $r \in [a, b]$ where $a := \inf_{x\in X} f(x) - \langle T^*\phi,\ x\rangle$ and $b := \sup_{z\in X} -g(Tz) - \langle T^*\phi,\ z\rangle$. One could then derive all the theorems using the sandwich theorem, in particular the Hahn–Banach theorem 3 and the subdifferential sum rule, Theorem 8, as consequences of Fenchel duality instead. This establishes the *Hahn–Banach/Fenchel duality* circle: Each of these facts is *equivalent* and easily interderivable with the nonemptiness of the subgradient of a function at a point of continuity.

An immediate consequence of Fenchel duality is a calculus of polar cones. Define the negative polar cone of a set K in a Banach space X by

$$K^- = \{x^* \in X^* \mid \langle x^*,\ x\rangle \le 0\ \forall\ x \in K\}. \tag{40}$$

An important example of a polar cone that has appeared in the applications discussed in this chapter is the *normal cone* of a convex set K at a point $x \in K$, defined by (33). Note that

$$N_K(\overline{x}) := (K - \overline{x})^-. \tag{41}$$

Corollary 3 (Polar Cone Calculus). *Let X and Y be Banach spaces and $K \subset X$ and $H \subset Y$ be cones, and let $A : X \to Y$ be a bounded linear map. Then*

$$K^- + A^*H^- \subset \left(K + A^{-1}H\right)^-$$

where equality holds if K and H are closed convex cones which satisfy $H - AK = Y$.

This can be used to easily establish the normal cone calculus for closed convex sets C_1 and C_2 at a point $x \in C_1 \cap C_2$

$$N_{C_1\cap C_2}(x) \supset N_{C_1}(x) + N_{C_2}(x)$$

with equality holding if, in addition, $0 \in \text{core}\,(C_1 - C_2)$ or $C_1 \cap \text{int}\, C_2 \neq \emptyset$.

Optimality Conditions Another important consequence of these ideas is the Pshenichnyi–Rockafellar [101, 103] condition for optimality for nonsmooth constrained optimization.

Theorem 13 (Pshenichnyi–Rockafellar Conditions). *Let X be a Banach space, let $C \subset X$ be closed and convex, and let $f : X \to (-\infty, +\infty]$ be a convex function. Suppose that either* $\operatorname{int} C \cap \operatorname{dom} f \neq \emptyset$ *and f is bounded below on C, or $C \cap \operatorname{cont} f \neq \emptyset$. Then there is an affine function $\alpha \leq f$ with* $\inf_C f = \inf_C \alpha$. *Moreover, $\overline{x}$ is a solution to*

$$(\mathcal{P}_0) \quad \begin{array}{ll} \underset{x \in X}{\text{minimize}} & f(x) \\ \text{subject to} & x \in C \end{array}$$

if and only if

$$0 \in \partial f(\overline{x}) + N_C(\overline{x}).$$

Guide. Apply the subdifferential sum rule to $f + \iota_C$ at $\overline{x}$. ■

A slight generalization extends this to linear constraints

$$(\mathcal{P}_{\text{lin}}) \quad \begin{array}{ll} \underset{x \in X}{\text{minimize}} & f(x) \\ \text{subject to} & Tx \in D \end{array}$$

Theorem 14 (First-Order Necessary and Sufficient). *Let X and Y be Banach spaces with $D \subset Y$ convex, and let $f : X \to (-\infty, +\infty]$ be convex and $T : X \to Y$ a bounded linear mapping. Suppose further that one of the following holds:*

$$0 \in \operatorname{core}(D - T \operatorname{dom} f), \ D \text{ is closed and } f \text{ is lsc}, \tag{42}$$

or

$$T \operatorname{dom} f \cap \operatorname{int}(D) \neq \emptyset. \tag{43}$$

Then the feasible set $C := \{x \in X \mid Tx \in D\}$ satisfies

$$\partial(f + \iota_C)(x) = \partial f(x) + T^*(N_D(Tx)) \tag{44}$$

and $\overline{x}$ is a solution to ($\mathcal{P}_{\text{lin}}$) if and only if

$$0 \in \partial f(\overline{x}) + T^*(N_D(T\overline{x})). \tag{45}$$

A point $y^* \in Y^*$ satisfying $T^* y^* \in -\partial f(\overline{x})$ in Theorem 14 is a *Lagrange multiplier*.

Lagrangian Duality The setting is limited to Euclidean space and the general convex program

$$(\mathcal{P}_{\text{cvx}}) \quad \begin{array}{ll} \underset{x\in E}{\text{minimize}} & f_0(x) \\ \text{subject to} & f_j(x) \le 0 \quad (j = 1, 2, \ldots, m) \end{array}$$

where the functions f_j for $j = 0, 1, 2, \ldots, m$ are convex and satisfy

$$\bigcap_{j=0}^{m} \text{dom } f_j \neq \emptyset. \tag{46}$$

Define the *Lagrangian* $L : E \times \mathbb{R}^m{}_+ \to (-\infty, +\infty]$ by

$$L(x, \lambda) := f_0(x) + \lambda^T F(x),$$

where $F := (f_1, f_2, \ldots, f_m)^T$. A *Lagrange multiplier* in this context is a vector $\overline{\lambda} \in \mathbb{R}^m{}_+$ for a feasible solution $\overline{x}$ if $\overline{x}$ minimizes the function $L(\cdot, \overline{\lambda})$ over E and $\overline{\lambda}$ satisfies the so-called *complimentary slackness conditions*: $\overline{\lambda}_j = 0$ whenever $f_j(\overline{x}) < 0$. On the other hand, if $\overline{x}$ is feasible for the convex program $(\mathcal{P}_{\text{cvx}})$ and there is a Lagrange multiplier, then $\overline{x}$ is optimal. Existence of the Lagrange multiplier is guaranteed by the following *Slater constraint qualification* first introduced in the 1950s.

Assumption 7 (Slater Constraint Qualification). *There exists an* $\hat{x} \in \text{dom } f_0$ *with* $f_j(\hat{x}) < 0$ *for* $j = 1, 2, \ldots, m$.

Theorem 15 (Lagrangian Necessary Conditions). *Suppose that* $\overline{x} \in \text{dom } f_0$ *is optimal for the convex program* $(\mathcal{P}_{cvx})$ *and that Assumption 7 holds. Then there is a Lagrange multiplier vector for* $\overline{x}$.

Guide. See [27, Theorem 3.2.8]. ■

Denote the optimal value of $(\mathcal{P}_{\text{cvx}})$ by p. Note that, since

$$\sup_{\lambda\in\mathbb{R}^m{}_+} L(x, \lambda) = \begin{cases} f(x) & \text{if } x \in \text{dom } f \\ +\infty & \text{otherwise,} \end{cases}$$

then

$$p = \inf_{x\in E} \sup_{\lambda\in\mathbb{R}^m{}_+} L(x, \lambda). \tag{47}$$

It is natural, then to consider the problem

$$d = \sup_{\lambda\in\mathbb{R}^m{}_+} \inf_{x\in E} L(x, \lambda) \tag{48}$$

where d is the *dual value*. It follows immediately that $p \geq d$. The difference between d and p is called the *duality gap*. The interesting problem is to determine when the gap is zero, that is, when $d = p$.

Theorem 16 (Dual Attainment). *If Assumption 7 holds for the convex programming problem ($\mathcal{P}_{cvx}$), then the primal and dual values are equal and the dual value is attained if finite.*

Guide. For a more detailed treatment of the theory of Lagrangian duality see [27, Sect. 4.3]. ■

Optimality and Lagrange Multipliers

In the previous sections, duality theory was presented as a byproduct of the Hahn–Banach/Fenchel duality circle of ideas. This provides many entry points to the theory of convex and variational analysis. For present purposes, however, the real significance of duality lies with its power to illuminate duality in convex optimization, not only as a theoretical phenomenon but also as an algorithmic strategy.

In order to get to optimality criteria and the existence of solutions to convex optimization problems, turn to the approximation of minima, or more generally the *regularity* and *well-posedness* of convex optimization problems. Due to its reliance on the Slater constraint qualification (49), Theorem 16 is not adequate for problems with equality constraints:

$$(\mathcal{P}_{eq}) \quad \begin{array}{ll} \underset{x \in S}{\text{minimize}} & f_0(x) \\ \text{subject to} & F(x) = 0 \end{array}$$

for $S \subset E$ closed and $F : E \to Y$ a Fréchet differentiable mapping between the Euclidean spaces E and Y.

More generally, one has problems of the form

$$(\mathcal{P}_E) \quad \begin{array}{ll} \underset{x \in S}{\text{minimize}} & f_0(x) \\ \text{subject to} & F(x) \in D \end{array} \tag{49}$$

for E and Y Euclidean spaces, and $S \subset E$ and $D \subset Y$ are convex but not necessarily with nonempty interior.

Example 6 (Simple Karush–Kuhn–Tucker). For linear optimization problems, relatively elementary linear algebra is all that is needed to assure the existence of Lagrange multipliers. Consider

$$(\mathcal{P}_E) \quad \begin{array}{ll} \underset{x\in S}{\text{minimize}} & f_0(x) \\ \text{subject to} & f_j(x) \in D_j, \quad j = 1, 2, \ldots, m \end{array}$$

for $f_j : \mathbb{R}^n \to \mathbb{R}$ $(j = 0, 1, 2, \ldots, s)$ continuously differentiable, $f_j : \mathbb{R}^n \to \mathbb{R}$ $(j = s+1, \ldots, m)$ linear. Suppose $S \subset E$ is closed and convex, while $D_i := (-\infty, 0]$ for $j = 1, 2, \ldots, s$ and $D_j := \{0\}$ for $j = s+1, \ldots, m$.

Theorem 17. *Denote by $f'_{\mathbb{J}}(x)$ the submatrix of the Jacobian of $(f_1, \ldots, f_s)^T$ (assuming this is defined at x) consisting only of those f'_j for which $f_j(x) = 0$. In other words, $f'_{\mathbb{J}}(x)$ is the Jacobian of the* active inequality constraints *at x. Let $\overline{x}$ be a local minimizer for $(\mathcal{P}_E)$ at which f_j are continuously differentiable $(j = 0, 1, \ldots, s)$ and the matrix*

$$\begin{pmatrix} f'_{\mathbb{J}}(\overline{x}) \\ A \end{pmatrix} \tag{50}$$

is full-rank where $A := (\nabla f_{s+1}, \ldots, \nabla f_m)^T$. Then there are $\overline{\lambda} \in \mathbb{R}^s$ and $\overline{\mu} \in \mathbb{R}^m$ satisfying

$$\overline{\lambda} \geq 0. \tag{51a}$$

$$(f_1(\overline{x}), \ldots, f_s(\overline{x}))\overline{\lambda} = 0. \tag{51b}$$

$$f'_0(\overline{x}) + \sum_{j=1}^{s} \overline{\lambda}_j f'_j(\overline{x}) + \overline{\mu}^T A = 0. \tag{51c}$$

Guide. An elegant and elementary proof is given in [36]. ■

For more general constraint structure, *regularity* of the feasible region is essential for the normal cone calculus which plays a key role in the requisite optimality criteria. More specifically, one has the following constraint qualification.

Assumption 8 (Basic Constraint Qualification).

$y = (0, \ldots, 0)$ is the only solution in $N_D(F(\overline{x}))$ to $0 \in \nabla F^T(\overline{x})y + N_S(\overline{x})$.

Theorem 18 (Optimality on Sets with Constraint Structure). *Let*

$$C = \{x \in S \mid F(x) \in D\}$$

for $F = (f_1, f_2, \ldots, f_m) : E \to \mathbb{R}^m$ with f_j continuously differentiable ($j = 1, 2, \ldots, m$), $S \subset E$ closed, and for $D = D_1 \times D_2 \times \cdots D_m \subset \mathbb{R}^m$ with D_j closed intervals ($j = 1, 2, \ldots, m$). Then for any $\overline{x} \in C$ at which Assumption 8 is satisfied one has

$$N_C(\overline{x}) = \nabla F^T(\overline{x}) N_D(F(\overline{x})) + N_S(\overline{x}). \tag{52}$$

If, in addition, f_0 is continuously differentiable and $\overline{x}$ is a locally optimal solution to $(\mathcal{P}_E)$ then there is a vector $\overline{y} \in N_D(F(\overline{x}))$, called a Lagrange multiplier *such that $0 \in \nabla f_0(\overline{x}) + \nabla F^T(\overline{x})\overline{y} + N_S(\overline{x})$.*

Guide. See [107, Theorems 6.14 and 6.15]. ■

Variational Principles

The Slater condition (49) is an *interiority* condition on the solutions to optimization problems. Interiority is just one type of *regularity* required of the solutions, wherein one is concerned with the behavior of solutions under perturbations. The next classical result lays the foundation for many modern notions of regularity of solutions.

Theorem 19 (Ekeland's Variational Principle). *Let (X, d) be a complete metric space and let $f : X \to (-\infty, +\infty]$ be a lsc function bounded from below. Suppose that $\epsilon > 0$ and $z \in X$ satisfy*

$$f(z) < \inf_X f + \epsilon.$$

For a given fixed $\lambda > 0$, there exists $y \in X$ such that

(i) $d(z, y) \leq \lambda$.
(ii) $f(y) + \frac{\epsilon}{\lambda} d(z, y) \leq f(z)$.
(iii) $f(x) + \frac{\epsilon}{\lambda} d(x, y) > f(y)$, *for all* $x \in X \backslash \{y\}$.

Guide. For a proof see [61]. For a version of the principle useful in the presence of symmetry see [32]. ■

An important application of Ekeland's variational principle is to the theory of subdifferentials. Given a function $f : X \to (-\infty, +\infty]$, a point $x_0 \in \text{dom } f$ and $\epsilon \geq 0$, the *ϵ-subdifferential* of f at x_0 is defined by

$$\partial_\epsilon f(x_0) = \{\phi \in X^* \mid \langle \phi, x - x_0 \rangle \leq f(x) - f(x_0) + \epsilon, \ \forall x \in X\}.$$

If $x_0 \notin \text{dom } f$ then by convention $\partial_\epsilon f(x_0) := \emptyset$. When $\epsilon = 0$ one has $\partial_\epsilon f(x) = \partial f(x)$. For $\epsilon > 0$ the domain of the ϵ-subdifferential coincides with dom f when f is a proper convex lsc function.

Theorem 20 (Brønsted–Rockafellar). *Suppose f is a proper lsc convex function on a Banach space X. Then given any $x_0 \in \text{dom } f, \epsilon > 0, \lambda > 0$ and $w_0 \in \partial_\epsilon f(x_0)$*

there exist $x \in \operatorname{dom} f$ *and* $w \in X^*$ *such that*

$$w \in \partial f(x), \quad \|x - x_0\| \leq \epsilon/\lambda \quad \textit{and} \quad \|w - w_0\| \leq \lambda.$$

In particular, the domain of ∂f *is dense in* $\operatorname{dom} f$.

Guide. Define $g(x) := f(x) - \langle w_0, x\rangle$ on X, a proper lsc convex function with the same domain as f. Then $g(x_0) \leq \inf_X g(x) + \epsilon$. Apply Theorem 19 to yield a nearby point y that is the minimum of a slightly perturbed function, $g(x)+\lambda\|x-y\|$. Define the new function $h(x) := \lambda\|x - y\| - g(y)$, so that $h(x) \leq g(x)$ for all X. The sandwich theorem (Theorem 7) establishes the existence of an affine separator $\alpha + \phi$ which is used to construct the desired element of $\partial f(x)$. ■

A nice application of Ekeland's variational principle provides an elegant proof of Klee's problem in Euclidean spaces [75]: Is every Čebyčev set C convex? Here, a *Čebyčev set* is one with the property that every point in the space has a unique best approximation in C. A famous result is as follows.

Theorem 21. *Every Čebyčev set in a Euclidean space is closed and convex.*

Guide. Since, for every finite dimensional Banach space with smooth norm, approximately convex sets are convex, it suffices to show that C is approximately convex, that is, for every closed ball disjoint from C there is another closed ball disjoint from C of arbitrarily large radius containing the first. This follows from the mean value Theorem 6 and Theorem 19. See [24, Theorem 3.5.2]. It is not known whether the same holds for Hilbert space. ■

Fixed Point Theory and Monotone Operators

Another application of Theorem 19 is Banach's fixed point theorem.

Theorem 22. *Let* (X, d) *be a complete metric space and let* $\phi : X \to X$. *Suppose there is a* $\gamma \in (0, 1)$ *such that* $d(\phi(x), \phi(y)) \leq \gamma d(x, y)$ *for all* $x, y \in X$. *Then there is a unique* fixed point $\overline{x} \in X$ *such that* $\phi(\overline{x}) = \overline{x}$.

Guide. Define $f(x) := d(x, \phi(x))$. Apply Theorem 19 to f with $\lambda = 1$ and $\epsilon = 1 - \gamma$. The fixed point $\overline{x}$ satisfies $f(x) + \epsilon d(x, \overline{x}) \geq f(\overline{x})$ for all $x \in X$. ■

The next theorem is a celebrated result in convex analysis concerning the *maximality* of lsc proper convex functions. A monotone operator T on X is *maximal* if $\operatorname{gph} T$ cannot be enlarged in $X \times X$ without destroying the monotonicity of T.

Theorem 23 (Maximal Monotonicity of Subdifferentials). *Let $f : X \to (-\infty, +\infty]$ be a lsc proper convex function on a Banach space. Then ∂f is maximal monotone.*

Guide. The result was first shown by Moreau for Hilbert spaces [95, Proposition 12.b], and shortly thereafter extended to Banach spaces by Rockafellar [102, 104]. For a modern infinite dimensional proof see [1, 24]. This result fails badly in incomplete normed spaces [24]. ■

Maximal monotonicity of subdifferentials of convex functions lies at the heart of the success of algorithms as this is equivalent to *firm nonexpansiveness* of the *resolvent* of the subdifferential $(I + \partial f)^{-1}$ [92]. An operator T is *firmly nonexpansive* on a closed convex subset $C \subset X$ when

$$\|Tx - Ty\|^2 \leq \langle x - y, Tx - Ty \rangle \quad \text{for all } x, y \in X. \tag{53}$$

T is just *nonexpansive* on the closed convex subset $C \subset X$ if

$$\|Tx - Ty\| \leq \|x - y\| \quad \text{for all } x, y \in C. \tag{54}$$

Clearly, all firmly nonexpansive operators are nonexpansive. One of the most longstanding questions in geometric fixed point theory is whether a nonexpansive self-map T of a closed bounded convex subset C of a reflexive space X must have a fixed point. This is known to hold in Hilbert space.

4 Case Studies

One can now collect the dividends from the analysis outlined above for problems of the form

$$\begin{array}{ll} \underset{x \in C \subset X}{\text{minimize}} & I_\varphi(x) \\ \text{subject to} & Ax \in D \end{array} \tag{55}$$

where X and Y are real Banach spaces with continuous duals X^* and Y^*, C and D are closed and convex, $A : X \to Y$ is a continuous linear operator, and the integral functional $I_\varphi(x) := \int_T \varphi(x(t))\mu(dt)$ is defined on some vector subspace $L_p(T, \mu)$ of X.

Linear Inverse Problems with Convex Constraints

Suppose X is a Hilbert space, $D = \{b\} \in \mathbb{R}^m$ and $\varphi(x) := \frac{1}{2}\|x\|^2$. To apply Fenchel duality, rewrite (12) using the indicator function

$$\begin{array}{ll} \underset{x\in X}{\text{minimize}} & \frac{1}{2}\|x\|^2 + \iota_C(x) \\ \text{subject to} & Ax = b. \end{array} \tag{56}$$

Note that the problem is posed on an infinite dimensional space, while the constraints (the measurements) are finite dimensional. Here Fenchel duality is used to transform an infinite dimensional problem into a finite dimensional problem. Let $F := \{x \in C \subset E \mid Ax = b\}$ and let G denote the extensible set in E consisting of all measurement vectors b for which F is nonempty. Potter and Arun show that the existence of $\overline{y} \in \mathbb{R}^m$ such that $b = AP_C A^*\overline{y}$ is guaranteed by the constraint qualification $b \in \text{ri } G$, where ri denotes the *relative interior* [100, Corollary 2]. This is a special case of Assumption 6, which here reduces to $b \in \text{int } A(C)$. Though at first glance the latter condition is more restrictive, it is no real loss of generality since, if it fails, one can restrict the problem to range(A) which is closed. Then it turns out that $b \in A\text{qri } C$, the image of the *quasi-relative interior* of C [27, Exercise 4.1.20]. Assuming this holds Fenchel duality, Theorem 12, yields the dual problem

$$\sup_{y\in\mathbb{R}^m} \langle b,\ y\rangle - \left(\tfrac{1}{2}\|\cdot\|^2 + \iota_C\right)^* (A^*y), \tag{57}$$

whose value is equivalent to the value of the primal problem. This is a finite dimensional unconstrained convex optimization problem whose solution is characterized by the inclusion (Theorem 4)

$$0 \in \partial\left(\tfrac{1}{2}\|\cdot\|^2 + \iota_C\right)^* (A^*y) - b. \tag{58}$$

Now from Lemma 2, Examples 5(ii) and (iii), and (37),

$$\left(\tfrac{1}{2}\|\cdot\|^2 + \iota_C\right)^* (x) = \left(\sigma_C \square \tfrac{1}{2}\|\cdot\|\right)(x) = \inf_{z\in X} \sigma_C(z) + \tfrac{1}{2}\|x - z\|^2.$$

The argmin of the Yosida approximation above (see Theorem 11) is the proximal operator (37). Applying the sum rule for differentials, Theorem 8 and Proposition 10 yield

$$\text{prox}_{1,\sigma_C}(x) = \text{argmin}_{z\in X} \left\{\sigma_C(z) + \tfrac{1}{2}\|z - x\|^2\right\} = x - P_C(x), \tag{59}$$

where P_C is the orthogonal projection onto the set C. This together with (58) yields the optimal solution $\overline{y}$ to (57):

$$b = AP_C(A^*\overline{y}). \tag{60}$$

Note that the existence of a solution to (60) is guaranteed by Assumption 6. This yields the solution to the primal problem as $\overline{x} = P_C(A^*\overline{y})$.

With the help of (59), the iteration proposed in [100] can be seen as a subgradient descent algorithm for solving

$$\inf_{y\in\mathbb{R}^m} h(y) := \sigma_C(A^*y - P_C(A^*y)) + \tfrac{1}{2}\|P_C(A^*y)\|^2 - \langle b,\ y\rangle\,.$$

The proposed algorithm is, given $y_0 \in \mathbb{R}^m$ generates the sequence $\{y_n\}_{n=0}^{\infty}$ by

$$y_{n+1} = y_n - \lambda\partial h(y_n) = y_n + \lambda\left(b - AP_C A^* y_n\right).$$

For convergence results of this algorithm in a much larger context see [55].

Imaging with Missing Data

This application is formally simpler than the previous example since there is no abstract constraint set. As discussed in section "Imaging with Missing Data" in Sect. 1 relaxations to the conventional problem take the form

$$\begin{array}{ll} \underset{x\in\mathbb{R}^n}{\text{minimize}} & I_{\varphi^*_{\epsilon,L}}(x) \\ \text{subject to} & Ax = b, \end{array} \tag{61}$$

where

$$\varphi^*_{\epsilon,L}(x) = \frac{\epsilon}{\ln(2)}\ln\left(4^{xL/\epsilon} + 1\right) - xL - \epsilon. \tag{62}$$

Using Fenchel duality, the dual to this problem is the concave optimization problem

$$\sup_{y\in\mathbb{R}^m} y^T b - I_{\varphi_{\epsilon,L}}(A^*y),$$

where

$$\varphi_{\epsilon,L}(x) := \epsilon\left(\frac{(L+x)\ln(L+x) + (L-x)\ln(L-x)}{2L\ln(2)} - \frac{\ln(L)}{\ln(2)}\right)$$
$$L, \epsilon > 0\, x \in [-L, L].$$

If there exists a point $\overline{y}$ satisfying $b = AA^*\overline{y}$, then the optimal value in the dual problem is attained and the primal solution is given by $A^*\overline{y}$. The objective in the dual problem is smooth and convex, so we could apply any number of efficient unconstrained optimization algorithms. Also, for this relaxation, the same numerical techniques can be used for all $L \to 0$. For further details see [28].

Inverse Scattering

Theorem 24. *Let X, Y be reflexive Banach spaces with duals X^* and Y^*. Let $F : Y^* \to Y$ and $G : X \to Y$ be bounded linear operators with $F = GSG^*$ for $S : X^* \to X$ a bounded linear operator satisfying the coercivity condition*

$$|\langle \varphi,\ S\varphi\rangle| \geq c\|\varphi\|_{X^*}^2 \quad \text{for some } c > 0 \text{ and all } \varphi \in \text{range}(G^*) \subset X^*.$$

Define $h(\psi) : Y^* \to (-\infty, +\infty] := |\langle \psi, F\psi\rangle|$, and let h^* denote the Fenchel conjugate of h. Then $\text{range}(G) = \text{dom}\ h^*$.

Proof. Following [74, Theorem 1.16], one shows that $h^*(\phi) = \infty$ for $\phi \notin \text{range}(G)$. To do this it is helpful to work with a dense subset of range G: $G^*(C)$ for $C := \{\psi \in Y^* \mid \langle \psi,\ \phi\rangle = 0\}$. It was shown in [74, Theorem 1.16] that $G^*(C)$ is dense in $\text{range}(G)$.

Now by the Hahn–Banach theorem 3 there is a $\hat{\phi} \in Y^*$ such that $\left\langle \hat{\phi},\ \phi \right\rangle = 1$. Since $G^*(C)$ is dense in $\text{range}(G^*)$, there is a sequence $\{\psi_n\}_{n=1}^{\infty} \subset C$ with

$$G^*\psi_n \to -G^*\hat{\phi}, \quad n \to \infty.$$

Now set $\psi_n := \hat{\psi}_n + \hat{\phi}$. Then $\langle \phi,\ \alpha\psi_n\rangle = \alpha$ and $G^*(\alpha\psi_n) = \alpha G^*\psi_n \to 0$ for any $\alpha \in \mathbb{R}$. Using the factorization of F one has

$$|\langle \psi_n,\ F\psi_n\rangle| = |\langle G^*\psi_n,\ SG^*\psi_n\rangle| \leq \|S\|\|G^*\psi_n\|_{X^*}^2$$

hence $\alpha^2\langle \psi_n,\ F\psi_n\rangle \to 0$ as $n \to \infty$ for all α, but $\langle \phi,\ \alpha\psi_n\rangle = \alpha$, that is, $\langle \phi, \alpha\psi_n\rangle - h(\alpha\psi_n) \to \alpha$ and $h^*(\phi) = \infty$. ∎

In the scattering application, one has a scatterer supported on a domain $D \subset \mathbb{R}^m$ ($m = 2$ or 3) that is illuminated by an incident field. The Helmholtz equation models the behavior of the fields on the exterior of the domain and the boundary data belongs to $X = H^{1/2}(\Gamma)$. On the sphere at infinity the leading-order behavior of the fields, the so-called far field pattern, lies in $Y = L^2(\mathbb{S})$. The operator mapping the boundary condition to the far field pattern – the *data-to-pattern operator* – is $G : H^{1/2}(\Gamma) \to L^2(\mathbb{S})$. Assume that the *far field operator* $F : L^2(\mathbb{S}) \to L^2(\mathbb{S})$ has the factorization $F = GS^*G^*$, where $S : H^{-1/2}(\Gamma) \to H^{1/2}(\Gamma)$ is a *single layer boundary operator* defined by

$$(S\varphi)(x) := \int_\Gamma \Phi(x, y)\varphi(y)ds(y),\ x \in \Gamma,$$

for $\Phi(x, y)$ the fundamental solution to the Helmholtz equation. With a few results about the denseness of G and the coercivity of S, which, though standard, will be glossed over, one has the following application to inverse scattering.

Corollary 4 (Application to Inverse Scattering). *Let $D \subset \mathbb{R}^m$ ($m = 2$ or 3) be an open bounded domain with connected exterior and boundary Γ. Let $G : H^{1/2}(\Gamma) \to L^2(\mathbb{S})$, be the* data-to-pattern *operator, $S : H^{-1/2}(\Gamma) \to H^{1/2}(\Gamma)$, the* single layer boundary operator *and let the* far field pattern *$F : L^2(\mathbb{S}) \to L^2(\mathbb{S})$ have the factorization $F = GS^*G^*$. Assume k^2 is not a Dirichlet eigenvalue of $-\triangle$ on D. Then* range $G =$ dom h^* *where $h(\psi) : L^2(\mathbb{S}) \to (-\infty, +\infty] := |\langle \psi, F\psi \rangle|$.*

Fredholm Integral Equations

The introduction featured the failure of Fenchel duality for Fredholm integral equations. The following sketch of a result on regularizations, or relaxations, recovers duality relationships. The result will show that by introducing a relaxation, one can recover the solution to ill-posed integral equations as the norm limit of solutions computable from a dual problem of maximum entropy type.

Theorem 25 ([22], Theorem 3.1). *Let $X = L_1(T, \mu)$ on a complete measure finite measure space and let $(Y, \|\cdot\|)$ be a normed space. The infimum $\inf_{x \in X} \{ I_\varphi(x) \mid Ax = b \}$ is attained when finite. In the case where it is finite, consider the relaxed problem for $\epsilon > 0$*

$$\left(\mathcal{P}^\epsilon_{MEP}\right) \quad \begin{array}{ll} \underset{x \in X}{\textit{minimize}} & I_\varphi(x) \\ \textit{subject to} & \|Ax - b\| \leq \epsilon. \end{array}$$

Let p_ϵ denote the value of $\left(\mathcal{P}^\epsilon_{MEP}\right)$. The value of p_ϵ equals d_ϵ, the value of the dual problem

$$\left(\mathcal{P}^\epsilon_{DEP}\right) \quad \underset{y^* \in Y^*}{\textit{maximize}} \ \langle b,\ y^* \rangle - \epsilon \|y^*\|_* - I_{\varphi^*}(A^* y^*),$$

and the unique optimal solution of $\left(\mathcal{P}^\epsilon_{MEP}\right)$ is given by

$$\overline{x}_{\varphi,\epsilon} := \frac{\partial \varphi^*}{\partial r}\left(A^* y^*_\epsilon\right),$$

where y^*_ϵ is any solution to $\left(\mathcal{P}^\epsilon_{DEP}\right)$. Moreover, as $\epsilon \to 0^+$, $\overline{x}_{\varphi,\epsilon}$ converges in mean to the unique solution of $\left(\mathcal{P}^0_{MEP}\right)$ and $p_\epsilon \to p_0$.

Guide. Attainment of the infimum in $\inf_{x \in X} \{ I_\varphi(x) \mid Ax = b \}$ follows from *strong convexity* of I_φ [26, 112]: strictly convex with weakly compact lower-level sets and with the *Kadec property*, i.e., that weak convergence together with convergence of the function values implies norm convergence. Let $g(y) := \iota_S(y)$ for $S = \{ y \in Y \mid b \in y + \epsilon B_Y \}$ and rewrite $\left(\mathcal{P}^\epsilon_{MEP}\right)$ as $\inf \{ I_\varphi(x) + g(Ax) \mid x \in X \}$. An elementary calculation shows that the Fenchel dual to $\left(\mathcal{P}^\epsilon_{MEP}\right)$ is $\left(\mathcal{P}^\epsilon_{DEP}\right)$. The

relaxed problem $\left(\mathcal{P}^{\epsilon}_{MEP}\right)$ has a constraint for which a Slater-type constraint qualification holds at any feasible point for the unrelaxed problem. The value d_ϵ is thus attained and equal to p_ϵ. Subgradient arguments following [25] show that $\overline{x}_{\varphi,\epsilon}$ is feasible for $\left(\mathcal{P}^{\epsilon}_{MEP}\right)$ and is the unique solution to $\left(\mathcal{P}^{\epsilon}_{MEP}\right)$. Convergence follows from weak compactness of the lower-level set $L(p_0) := \left\{x \;\middle|\; I_\varphi(x) \leq p_0\right\}$, which contains the sequence $(\overline{x}_{\varphi,\epsilon})_{\epsilon>0}$. Weak convergence of $\overline{x}_{\varphi,\epsilon}$ to the unique solution to the unrelaxed problem follows from strict convexity of I_φ. Convergence of the function values and strong convexity of I_φ then yields norm convergence. ■

Notice that the dual in Theorem 25 is unconstrained and easier to compute with, especially when there are finitely many constraints. This theorem remains valid for objectives of the form $I_\varphi(x) + \langle x*,\ x\rangle$ for $x*$ in $L_\infty(T)$. This enables one to apply them to many *Bregman distances*, that is, integrands of the form $\phi(x) - \phi(x_0) - \langle \phi'(x_0),\ x - x_0\rangle$, where ϕ is closed and convex on $\mathbb{R}$.

5 Open Questions

Regrettably, due to space constraints, fixed point theory and many facts about monotone operators that are useful in proving convergence of algorithms have been omitted. However, it is worthwhile noting two long-standing problems that impinge on fixed point and monotone operator theory.

1. Klee's problem: is every Čebyčev set C in a Hilbert space convex?
2. Must a nonexpansive self-map T of a closed bounded convex subset C of a reflexive space X have a fixed point?

6 Conclusion

Duality and convex programming provides powerful techniques for solving a wide range of imaging problems. While frequently a means toward computational ends, the dual perspective can also yield new insight into image processing problems and the information content of data implicit in certain models. Five main applications illustrate the convex analytical approach to problem solving and the use of duality: linear inverse problems with convex constraints, compressive imaging, image denoising and deconvolution, nonlinear inverse scattering, and finally Fredholm integral equations. These are certainly not exhaustive, but serve as good templates. The Hahn–Banach/Fenchel duality cycle of ideas developed here not only provides a variety of entry points into convex and variational analysis, but also underscores duality in convex optimization as both a theoretical phenomenon and an algorithmic strategy.

As readers of this volume will recognize, not all problems of interest are convex. But just as nonlinear problems are approached numerically by sequences of linear

approximations, nonconvex problems can be approached by sequences of convex approximations. Convexity is the central organizing principle and has tremendous algorithmic implications, including not only computable guarantees about solutions, but efficient means toward that end. In particular, convexity implies the existence of implementable, polynomial-time, algorithms. This chapter is meant to be a foundation for more sophisticated methodologies applied to more complicated problems.

Acknowledgments D. Russell Luke's work was supported in part by NSF grants DMS-0712796 and DMS-0852454. Work on the second edition was supported by DFG grant SFB755TPC2. The authors wish to thank Matthew Tam for his assistance in preparing a revision for the second edition of the handbook.

Cross-References

▸ Compressive Sensing
▸ Inverse Scattering
▸ Iterative Solution Methods
▸ Numerical Methods and Applications in Total Variation Image Restoration
▸ Regularization Methods for Ill-Posed Problems
▸ Total Variation in Imaging
▸ Variational Methods in Shape Analysis

References

1. Alves, M., Svaiter, B.F.: A new proof for maximal monotonicity of subdifferential operators. J. Convex Anal. **15**(2), 345–348 (2008)
2. Aragón Artacho, F.J., Borwein, J.M.: Global convergence of a non-convex Douglas–Rachford iteration. J. Glob. Optim. **57**(3), 753–769 (2013)
3. Aubert, G., Kornprost, P.: Mathematical Problems Image Processing. Applied Mathematical Sciences, vol. 147. Springer, New York (2002)
4. Auslender, A., Teboulle, M.: Asymptotic Cones and Functions in Optimization and Variational Inequalities. Springer, New York (2003)
5. Bauschke, H.H., Borwein, J.M.: On projection algorithms for solving convex feasibility problems. SIAM Rev. **38**(3), 367–426 (1996)
6. Bauschke, H.H., Combettes, P.L.: Convex Analysis and Monotone Operator Theory in Hilbert Spaces. CMS Books in Mathematics. Springer, New York (2011)
7. Bauschke, H.H., Combettes, P.L., Luke, D.R.: Phase retrieval, error reduction algorithm and Fienup variants: a view from convex feasibility. J. Opt. Soc. Am. A **19**(7), 1334–1345 (2002)
8. Bauschke, H.H., Combettes, P.L., Luke, D.R.: A hybrid projection reflection method for phase retrieval. J. Opt. Soc. Am. A **20**(6), 1025–1034 (2003)
9. Bauschke, H.H., Combettes, P.L., Luke, D.R.: Finding best approximation pairs relative to two closed convex sets in Hilbert spaces. J. Approx. Theory **127**, 178–192 (2004)
10. Bauschke, H.H., Cruz, J.Y., Phan, H.M., Wang, X.: The rate of linear convergence of the Douglas–Rachford algorithm for subspaces is the cosine of the Friedrichs angle (2013). Preprint. arXiv:1309.4709v1 [math.OC]

11. Bauschke, H.H., Luke, D.R., Phan, H.M., Wang, X.: Restricted normal cones and the method of alternating projections: theory. Set-Valued Var. Anal. **21**, 431–473 (2013)
12. Bauschke, H.H., Luke, D.R., Phan, H.M., Wang, X.: Restricted normal cones and the method of alternating projections: applications. Set-Valued Var. Anal. **21**, 475–501 (2013)
13. Bauschke, H.H., Luke, D.R., Phan, H.M., Wang, X.: Restricted normal cones and sparsity optimization with affine constraints. Found Comput. Math. **14**(1), 63–83 (2014)
14. Bauscke, H.H., Phan, H.M., Wang, X.: The method of alternating relaxed projections for two nonconvex sets. Vietnam J. Math. (in press). doi:10.1007/510013-013-0049-8
15. Beck, A., Eldar, Y.: Sparsity constrained nonlinear optimization: optimality conditions and algorithms. SIAM J. Optim. **23**, 1480–1509 (2013)
16. Beck, A., Teboulle, M.: A linearly convergent algorithm for solving a class of nonconvex/affine feasibility problems. In: Bauschke, H.H., Burachik, R.S., Combettes, P.L., Elser, V., Luke, D.R., Wolkowicz, H. (eds.) Fixed-Point Algorithms for Inverse Problems in Science and Engineering. Springer Optimization and its Applications, pp. 33–48. Springer, New York (2011)
17. Bect, J., Blanc-Féraud, L., Aubert, G., Chambolle, A.: A ℓ_1-unified variational framework for image restoration. In: Pajdla, T., Matas, J. (eds.) Proceedings of the Eighth European Conference on Computer Vision, Prague, 2004. Lecture Notes in Computer Science, vol. 3024, pp. 1–13. Springer, New York (2004)
18. Ben-Tal, A., Borwein, J.M., Teboulle, M.: A dual approach to multidimensional l_p spectral estimation problems. SIAM J. Contr. Optim. **26**, 985–996 (1988)
19. Blumensath, T., Davies, M.E.: Iterative hard thresholding for compressed sensing. Appl. Comput. Harm. Anal. **27**, 265–274 (2009)
20. Blumensath, T., Davies, M.E.: Normalised iterative hard thresholding: guaranteed stability and performance. IEEE J. Sel. Top. Sign. Process. **4**, 298–309 (2010)
21. Bonnans, J.F., Gilbert, J.C., Lemaréchal, C., Sagastizábal, C.A.: Numerical Optimization, 2nd edn. Springer, New York (2006)
22. Borwein, J.M.: On the failure of maximum entropy reconstruction for Fredholm equations and other infinite systems. Math. Program **61**, 251–261 (1993)
23. Borwein, J.M., Hamilton, C.: Symbolic fenchel conjugation. Math. Program **116**, 17–35 (2009)
24. Borwein, J.M., Jon Vanderwerff, J.: Convex Functions: Constructions, Characterizations and Counterexamples. Encyclopedias in Mathematics, vol. 109. Cambridge University Press, New York (2009)
25. Borwein, J.M., Lewis, A.S.: Duality relationships for entropy-like minimization problems. SIAM J. Contr. Optim. **29**, 325–338 (1990)
26. Borwein, J.M., Lewis, A.S.: Convergence of best entropy estimates. SIAM J. Optim. **1**, 191–205 (1991)
27. Borwein, J.M., Lewis, A.S.: Convex Analysis and Nonlinear Optimization: Theory and Examples, 2nd edn. Springer, New York (2006)
28. Borwein, J.M., Luke, D.R.: Entropic regularization of the ℓ_0 function. In: Bauschke, H.H., Burachik, R.S., Combettes, P.L., Elser, V., Luke, D.R., Wolowicz, H. (eds.) Fixed-Point Algorithms for Inverse Problems in Science and Engineering. Springer Optimization and its Applications, vol. 49, pp. 65–92. Springer, Berlin (2011)
29. Borwein, J.M., Sims, B.: The Douglas–Rachford algorithm in the absence of convexity. In: Bauschke, H.H., Burachik, R.S., Combettes, P.L., Elser, V., Luke, D.R., Wolowicz, H. (eds.) Fixed-Point Algorithms for Inverse Problems in Science and Engineering. Springer Optimization and its Applications, vol. 49, pp. 93–109. Springer, Berlin (2011)
30. Borwein, J.M., Tam, M.K.: A cyclic Douglas–Rachford iteration scheme. J. Optim. Theory Appl. (2013). doi:10.1007/s10957-013-0381-x
31. Borwein, J.M., Zhu, Q.J.: Techniques of Variational Analysis. CMS Books in Mathematics. Springer, New York (2005)
32. Borwein, J.M., Zhu, Q.J.: Variational methods in the presence of symmetry. Adv. Nonlinear Anal. **2**(3), 271–307 (2013)

33. Borwein, J.M., Lewis, A.S., Limber, M.N., Noll, D.: Maximum entropy spectral analysis using first order information. Part 2: a numerical algorithm for fisher information duality. Numer. Math. **69**, 243–256 (1995)
34. Borwein, J.M., Lewis, A.S., Noll, D.: Maximum entropy spectral analysis using first order information. Part 1: fisher information and convex duality. Math. Oper. Res. **21**, 442–468 (1996)
35. Boyd, S., Vandenberghe, L.: Convex Optimization. Oxford University Press, New York (2003)
36. Brezhneva, O.A., Tret'yakov, A.A., Wright, S.E.: A simple and elementary proof of the Karush–Kuhn–Tucker theorem for inequality-constrained optimization. Optim. Lett. **3**, 7–10 (2009)
37. Burg, J.P.: Maximum entropy spectral analysis. Paper presented at the 37th Meeting of the Society of Exploration Geophysicists, Oklahoma City (1967)
38. Burke, J.V., Luke, D.R.: Variational analysis applied to the problem of optical phase retrieval. SIAM J. Contr. Optim. **42**(2), 576–595 (2003)
39. Byrne, C.L.: Signal Processing: A Mathematical Approach. AK Peters, Natick (2005)
40. Candés, E., Tao, T.: Decoding by linear programming. IEEE Trans. Inform. Theory **51**(12), 4203–4215 (2005)
41. Candés, E., Tao, T.: Near-optimal signal recovery from random projections: universal encoding strategies? IEEE Trans. Inform. Theory **52**(12), 5406–5425 (2006)
42. Censor, Y., Zenios, S.A.: Parallel Optimization: Theory Algorithms and Applications. Oxford University Press, Oxford (1997)
43. Chambolle, A.: An algorithm for total variation minimization and applications. J Math. Imaging Vis. **20**, 89–97 (2004)
44. Chambolle, A., Lions, P.L.: Image recovery via total variation minimization and related problems. Numer. Math. **76**, 167–188 (1997)
45. Chambolle, A., Pock, T.: A first-order primal–dual algorithm for convex problems with applications to imaging. J. Math. Imaging Vis. **40**(1), 120–145 (2011)
46. Chan, T.F., Golub, G.H., Mulet, P.: A nonlinear primal–dual method for total variation based image restoration. SIAM J. Sci. Comput. **20**(6), 1964–1977 (1999)
47. Chen, S.S., Donoho, D.L., Saunders, M.A.: Atomic decomposition by basis pursuit. SIAM J. Sci. Comput. **20**(1), 33–61 (1999)
48. Chlamtac, E., Tulsiani, M.: Convex relaxations and integrality gaps. In: Anjos, M.F., Lasserre, J.B. (eds.) Handbook on Semidefinite, Convex and Polynomial Optimization. International Series in Operations Research & Management Science, vol. 166, pp. 139–169. Springer, New York (2012)
49. Clarke, F.H.: Optimization and Nonsmooth Analysis. Classics in Applied Mathematics, vol. 5. SIAM, Philadelphia (1990)
50. Clarke, F.H., Stern, R.J., Ledyaev, Yu.S., Wolenski, P.R.: Nonsmooth Analysis and Control Theory. Springer, New York (1998)
51. Combettes, P.L.: The convex feasibility problem in image recovery. In: Hawkes, P.W. (ed.) Advances in Imaging and Electron Physics, vol. 95, pp 155–270. Academic, New York (1996)
52. Combettes, P.L., Pesquet, J.-C.: Proximal splitting method in signal processing. In: Bauschke, H.H., Burachik, R.S., Combettes, P.L., Elser, V., Luke, D.R., Wolowicz, H. (eds.) Fixed-Point Algorithms for Inverse Problems in Science and Engineering. Springer Optimization and its Applications, vol. 49, pp. 185–212. Springer, Berlin (2011)
53. Combettes, P.L., Trussell, H.J.: Method of successive projections for finding a common point of sets in metric spaces. J. Optim. Theory Appl. **67**(3), 487–507 (1990)
54. Combettes, P.L., Wajs, V.R.: Signal recovery by proximal forward–backward splitting. SIAM J. Multiscale Model. Simulat. **4**(4), 1168–1200 (2005)
55. Combettes, P.L., Dũng, D., Vũ, B.C.: Dualization of signal recovery problems. Set-Valued Var. Anal. **18**, 373–404 (2010)
56. Dacunha-Castelle, D., Gamboa, F.: Maximum d'entropie et probléme des moments. l'Institut Henri Poincaré **26**, 567–596 (1990)

57. Destuynder, P., Jaoua, M., Sellami, H.: A dual algorithm for denoising and preserving edges in image processing. J. Inverse Ill-Posed Prob. **15**, 149–165 (2007)
58. Deutsch, F.: Best Approximation in Inner Product Spaces. CMS Books in Mathematics. Springer, New York (2001)
59. Donoho, D.L., Johnstone, I.M.: Ideal spatial adaptation by wavelet shrinkage. Biometrika **81**(3), 425–455 (1994)
60. Eggermont, P.P.B.: Maximum entropy regularization for Fredholm integral equations of the first kind. SIAM J. Math. Anal. **24**(6), 1557–1576 (1993)
61. Ekeland, I., Temam, R.: Convex Analysis and Variational Problems. Elsevier, New York (1976)
62. Fenchel, W.: On conjugate convex functions. Can. J. Math. **1**, 73–77 (1949)
63. Foucart, S.: Hard thresholding pursuit: an algorithm for compressive sensing. SIAM J. Numer. Anal. **49**(6), 2543–2563 (2011)
64. Goodrich, R.K., Steinhardt, A.: L_2 spectral estimation. SIAM J. Appl. Math. **46**, 417–428 (1986)
65. Grant, M., Boyd, S.: Graph implementations for nonsmooth convex programs. In: Blondel, V.D., Boyd, S.P., Kimura, H. (eds.) Recent Advances in Learning and Control. LNCIS 371, pp. 96–110. Springer-Verlag, Heidelberg (2008)
66. Groetsch, C.W.: The Theory of Tikhonov Regularization for Fredholm Integral Equations of the First Kind. Pitman, Bostan (1984)
67. Groetsch, C.W.: Stable Approximate Evaluation of Unbounded Operators. Lecture Notes in Mathematics, vol. 1894. Springer, New York (2007)
68. Hesse, R., Luke, D.R.: Nonconvex notions of regularity and convergence of fundamental algorithms for feasibility problems. SIAM J. Optim. **23**(4), 2397–2419 (2013). Preprint. arXiv:1212.3349v2 [math.OC]
69. Hesse, R., Luke, D.R., Neumann, P.: Projection Methods for Sparse Affine Feasibility: Results and Counterexamples (2013). Preprint. arXiv:1307.2009 [math.OC]
70. Hintermüller, M., Stadler, G.: An infeasible primal-dual algorithm for total bounded variation-based inf-convolution-type image restoration. SIAM J. Sci. Comput. **28**, 1–23 (2006)
71. Hiriart-Urruty, J.-B., Lemaréchal, C.: Convex Analysis and Minimization Algorithms, I and II. Grundlehren der mathematischen Wissenschaften, vols. 305–306. Springer, New York (1993)
72. Hiriart-Urruty, J.-B., Lemaréchal, C.: Fundamentals of Convex Analysis. Grundlehren der mathematischen Wissenschaften. Springer, New York (2001)
73. Iusem, A.N., Teboulle, M.: A regularized dual-based iterative method for a class of image reconstruction problems. Inverse Probl. **9**, 679–696 (1993)
74. Kirsch, A., Grinberg, N.: The Factorization Method for Inverse Problems. Oxford Lecture Series in Mathematics and its Applications, vol. 36. Oxford University Press, New York (2008)
75. Klee, V.: Convexity of Cebysev sets. Math. Ann. **142**, 291–304 (1961)
76. Kress, R.: Linear Integral Equations. Applied Mathematical Sciences, vol. 82, 2nd edn. Springer, New York (1999)
77. Levi, L.: Fitting a bandlimited signal to given points. IEEE Trans. Inform. Theory **11**, 372–376 (1965)
78. Lewis, A.S., Malick, J.: Alternating projections on manifolds. Math. Oper. Res. **33**, 216–234 (2008)
79. Lewis, A.S., Luke, D.R., Malick, J.: Local linear convergence of alternating and averaged projections. Found. Comput. Math. **9**(4), 485–513 (2009)
80. Lucchetti, R.: Convexity and Well-Posed Problems. CMS Books in Mathematics, vol. 22. Springer, New York (2006)
81. Lucet, Y.: Faster than the fast Legendre transform, the linear-time Legendre transform. Numer. Alg. **16**(2), 171–185 (1997)
82. Lucet, Y.: Fast Moreau envelope computation I: numerical algorithms. Numer. Alg. **43**(3), 235–249 (2006)

83. Lucet, Y.: Hybrid symbolic-numeric algorithms for computational convex analysis. Proc. Appl. Math. Mech. **7**(1), 1062301–1062302 (2007)
84. Lucet, Y.: What shape is your conjugate? A survey of computational convex analysis and its applications. SIAM J. Optim. **20**(1), 216–250 (2009)
85. Lucet, Y., Bauschke, H.H., Trienis, M.: The piecewise linear quadratic model for computational convex anlysis. Comput. Optim. Appl. **43**(1), 95–11 (2009)
86. Luenberger, D.G., Ye, Y.: Linear and Nonlinear Programming, 3rd edn. Springer, New York (2008)
87. Luke, D.R.: Relaxed averaged alternating reflections for diffraction imaging. Inverse Probl. **21**, 37–50 (2005)
88. Luke, D.R.: Finding best approximation pairs relative to a convex and a prox-regular set in Hilbert space. SIAM J. Optim. **19**(2), 714–739 (2008)
89. Luke, D.R., Burke, J.V., Lyon, R.G.: Optical wavefront reconstruction: theory and numerical methods. SIAM Rev. **44**, 169–224 (2002)
90. Maréchal, P., Lannes, A. (1997) Unification of some deterministic and probabilistic methods for the solution of inverse problems via the principle of maximum entropy on the mean. Inverse Probl. **13**, 135–151 (1962)
91. Mattingley, J., Body, S.: CVXGEN: a code generator for embedded convex optimization. Optim. Eng. **13**, 1–27 (2012)
92. Minty, G.J.: Monotone (nonlinear) operators in Hilbert space. Duke Math. J. **29**(3), 341–346 (1962)
93. Mordukhovich, B.S.: Variational Analysis and Generalized Differentiation, I: Basic Theory; II: Applications. Grundlehren der mathematischen Wissenschaften. Springer, New York (2006)
94. Moreau, J.J.: Fonctions convexes duales et points proximaux dans un espace Hilbertien. Comptes Rendus de l'Académie des Sciences de Paris **255**, 2897–2899 (1962)
95. Moreau, J.J.: Proximité et dualité dans un espace Hilbertian. Bull de la Soc math de France **93**(3), 273–299 (1965)
96. Nesterov, Y.E., Nemirovskii, A.S.: Interior-Point Polynomial Algorithms in Convex Programming. SIAM, Philadelphia (1994)
97. Nocedal, J., Wright, S.: Numerical Optimization. Springer, New York (2000)
98. Patrinos, P., Sarimveis, H.: Convex parametric piecewise quadratic optimization: theory and algorithms. Automatica **47**, 1770–1777 (2011)
99. Phelps, R.R.: Convex Functions, Monotone Operators and Differentiability. Lecture Notes in Mathematics, vol. 1364, 2nd edn. Springer, New York (1993)
100. Potter, L.C., Arun, K.S.: A dual approach to linear inverse problems with convex constraints. SIAM J. Contr. Opt. **31**(4), 1080–1092 (1993)
101. Pshenichnyi, B.N.: Necessary Conditions for an Extremum. Pure and Applied Mathematics, vol. 4. Marcel Dekker, New York (1971). Translated from Russian by Karol Makowski. Translation edited by Lucien W. Neustadt
102. Rockafellar, R.T.: Characterization of the subdifferentials of convex functions. Pacific J. Math. **17**, 497–510 (1966)
103. Rockafellar, R.T.: Convex Analysis. Princeton University Press, Princeton (1970)
104. Rockafellar, R.T.: On the maximal monotonicity of subdifferential mappings. Pacific J. Math. **33**, 209–216 (1970)
105. Rockafellar, R.T.: Integrals which are convex functionals, II. Pacific J. Math. **39**, 439–469 (1971)
106. Rockafellar, R.T.: Conjugate Duality and Optimization. SIAM, Philadelphia (1974)
107. Rockafellar, R.T., Wets, R.J.: Variational Analysis. Grundlehren der mathematischen Wissenschaften. Springer, Berlin (1998)
108. Rudin, L.I., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. Physica D **60**, 259–268 (1992)
109. Scherzer, O., Grasmair, M., Grossauer, H., Haltmeier, M., Lenzen, F.: Variational Methods in Imaging. Applied Mathematical Sciences, vol. 167. Springer, New York (2009)

110. Simons, S.: From Hahn–Banach to Monotonicity. Lecture Notes in Mathematics, vol. 1693. Springer, New York (2008)
111. Singer, I.: Duality for Nonconvex Approximation and Optimization. Springer, New York (2006)
112. Teboulle, M., Vajda, I.: Convergence of best φ-entropy estimates. IEEE Trans. Inform. Process. **39**, 279–301 (1993)
113. Tihonov, A.N.: On the regularization of ill-posed problems (Russian). Dokl. Akad. Nauk. SSSR **153**, 49–52 (1963)
114. Tropp, J.A.: Algorithms for simultaneous sparse approximation. Part II: convex relaxation. Signal Process. **86**(3), 589–602 (2006)
115. Tropp, J.A.: Just relax: convex programming methods for identifying sparse signals in noise. IEEE Trans. Inf. Theory **52**(3), 1030–1051 (2006)
116. Weiss, P., Aubert, G., Blanc-Féraud, L.: Efficient schemes for total variation minimization under constraints in image processing. SIAM J. Sci. Comput. **31**, 2047–2080 (2009)
117. Wright, S.J.: Primal–Dual Interior-Point Methods. SIAM, Philadelphia (1997)
118. Zălinescu, C.: Convex Analysis in General Vector Spaces. World Scientific, River Edge (2002)
119. Zarantonello, E.H.: Projections on convex sets in Hilbert space and spectral theory. In: Zarantonello, E.H. (ed.) Contributions to Nonlinear Functional Analysis, pp 237–424. Academic, New York (1971)

EM Algorithms

Charles Byrne and Paul P.B. Eggermont

Contents

1 Maximum Likelihood Estimation 306
2 The Kullback–Leibler Divergence 309
3 The EM Algorithm 310
The Maximum Likelihood Problem 310
The Bare-Bones EM Algorithm 311
The Bare-Bones EM Algorithm Fleshed Out 313
The EM Algorithm Increases the Likelihood 315
4 The EM Algorithm in Simple Cases 316
Mixtures of Known Densities 317
A Deconvolution Problem 319
The Deconvolution Problem with Binning 323
Finite Mixtures of Unknown Distributions 327
Empirical Bayes Estimation 329
5 Emission Tomography 330
Flavors of Emission Tomography 330
The Emission Tomography Experiment 331
The Shepp–Vardi EM Algorithm for PET 332
Prehistory of the Shepp–Vardi EM Algorithm 336
6 Electron Microscopy 337
Imaging Macromolecular Assemblies 337
The Maximum Likelihood Problem 337
The EM Algorithm, up to a Point 339
The Ill-Posed Weighted Least-Squares Problem 342
7 Regularization in Emission Tomography 342
The Need for Regularization 342

C. Byrne (✉)
Department of Mathematical Sciences, University of Massachusetts Lowell, Lowell, MA, USA
e-mail: Charles_Byrne@uml.edu

P.P.B. Eggermont
Food and Resource Economics, University of Delaware, Newark, DE, USA
e-mail: eggermon@udel.edu

© Springer Science+Business Media New York 2015

O. Scherzer (ed.), *Handbook of Mathematical Methods in Imaging*,
DOI 10.1007/978-1-4939-0790-8_8

Smoothed EM Algorithms 343
Good's Roughness Penalization 345
Gibbs Smoothing 347
8 Convergence of EM Algorithms 348
The Two Monotonicity Properties 349
Monotonicity of the Shepp–Vardi EM Algorithm 350
Monotonicity for Mixtures 352
Monotonicity of the Smoothed EM Algorithm 354
Monotonicity for Exact Gibbs Smoothing 359
9 EM-Like Algorithms 362
Minimum Cross-Entropy Problems 363
Nonnegative Least Squares 366
Multiplicative Iterative Algorithms 369
10 Accelerating the EM Algorithm 371
The Ordered Subset EM Algorithm 371
The ART and Cimmino–Landweber Methods 374
The MART and SMART Methods 377
Row-Action and Block-Iterative EM Algorithms 380
References 384

1 Maximum Likelihood Estimation

Expectation-maximization algorithms, or EM algorithms for short, are iterative algorithms designed to solve maximum likelihood estimation problems. The general setting is that one observes a random sample $Y_1, Y_2, \ldots, Y_n$ of a random variable Y whose probability density function (pdf) $f(\cdot \mid x_o)$ with respect to some (known) dominating measure is known up to an unknown "parameter" x_o. The goal is to estimate x_o and, one might add, to do it well. In this chapter, that means to solve the maximum likelihood problem

$$\text{maximize} \quad \prod_{i=1}^{n} f(Y_i \mid x) \quad \text{over} \quad x, \tag{1}$$

and to solve it by means of EM algorithms. The solution, assuming it exists and is unique, is called the *maximum likelihood estimator* of x_o. Here, the estimator is typically denoted by $\hat{x}$.

The notion of EM algorithms was coined by [27], who unified various earlier instances of EM algorithms and in particular emphasized the notion of "missing" data in maximum likelihood estimation problems, following Hartley [53]. Here, the missing data refers to data that were not observed. Although this seems to imply that these data could have or should have been observed, it is usually the case that these missing data are inherently inaccessible. Typical examples of this are deconvolution problems, but it may be instructive to describe a simplified version in the form of finite mixtures of probability densities.

Let the random variable Y be a mixture of some other continuous random variables $Z_1, Z_2, \ldots, Z_m$ for some known integer m. For each j, $1 \leqslant j \leqslant m$, denote the pdf of Z_j by $f(\cdot \mid j)$. The pdf of Y is then

$$f_Y(y) = \sum_{j=1}^{m} w_j^* \, f(y \mid j), \tag{2}$$

where $w^* = \left(w_1^*, w_2^*, \ldots, w_m^*\right)$ is a probability vector. In other words, the w_j^* are nonnegative and add up to 1. The interpretation is that for each j,

$$Y = Z_j \quad \text{with probability} \quad w_j^*. \tag{3}$$

As before, given a random sample $Y_1, Y_2, \ldots, Y_n$ of the random variable Y, the goal is to estimate the "parameter" w^*. The maximum likelihood problem for doing this is

$$\begin{aligned} &\text{maximize} \quad \prod_{i=1}^{m} \left\{ \sum_{j=1}^{m} w_j \, f(Y_i \mid j) \right\} \\ &\text{subject to} \quad w = (w_1, w_2, \ldots, w_m) \quad \text{is a probability vector.} \end{aligned} \tag{4}$$

Now, what are the "missing" data for this finite mixture problem? In view of the interpretation (3), it would clearly be useful if for each Y_i, it was known which random variable Z_j it was supposed to be a random sample of. Thus, let $J_i \in \{1, 2, \ldots, m\}$ such that

$$Y_i \mid J_i = Z_{J_i}. \tag{5}$$

Then, $J_1, J_2, \ldots, J_n$ would be a random sample of the random variable J, whose distribution would then be easy to estimate: for each j,

$$\hat{w}_j = \frac{\#\{J_i = j\}}{n}, \tag{6}$$

the fraction of the J_i that were equal to j. Note that the distribution of J is given by

$$\mathbb{P}[J = j] = w_j^*, \quad j = 1, 2, \ldots, m. \tag{7}$$

Of course, unfortunately, the J_i are not known. It is not even apparent that it would be advantageous to think about the J_i, but in fact it is, as this chapter tries to make clear.

From the image processing point of view, the above problem becomes more interesting if the finite sum in (2) is interpreted as a discretization of the integral transform

$$f_Y(y) = \int f(y \mid x) \, w(x) \, dx \tag{8}$$

and the goal is to recover the function or image w from the random sample $Y_1, Y_2, \ldots, Y_n$. The maximum likelihood problem of estimating w,

$$\text{maximize} \quad \prod_{i=1}^{n} \left\{ \int f(Y_i \mid x)\, w(x)\, dx \right\} \quad \text{over} \quad w, \tag{9}$$

is (formally) a straightforward extension of the mixture problem. Such (one- and two-dimensional) deconvolution problems abound in practice, e.g., in astronomy and tomography. See the suggested reading list.

In the next two sections, the bare essentials of a more-or-less general version of the EM algorithm are presented and it is shown that it increases the likelihood. Without special conditions, that is, all one can say about the convergence of the EM algorithm, one cannot even claim that in the limit, it achieves the maximum value of the likelihood. See [98]. For the convex case, where the negative log-likelihood is convex and the constraint set is convex as well, one can say much more, as will become clear.

Before discussing the two "big" applications of positron emission tomography (PET) and three-dimensional electron microscopy (3D-EM. Yes, another instance of EM!), it is prudent to discuss some simple examples of maximum likelihood estimation and to derive the associated EM algorithms. It turns that the prototypical example is that of estimating the weights in a mixture of known distributions; see (2). By analogy, this example shows how one should derive the EM algorithm for deconvolution problems with binned data, which is similar to the situation in positron emission tomography. The general parametric maximum likelihood estimation is also discussed, as well as the related case of empirical Bayes estimation. The latter has some similarity with 3D-EM.

All of this naturally leads to the discussion of the maximum likelihood approach to positron emission tomography (which originated with Rockmore and Macovski [82], but who mistakenly took the road of a least-squares treatment) and the EM algorithm of Shepp and Vardi [88]. This is one of the classic examples of Poisson data. However, even Poisson data may be interpreted as a random sample of some random variable; see section "The Emission Tomography Experiment." For the ubiquitous nature of Poisson data, see [4] and references therein.

The very messy example of the reconstruction of the shapes of macromolecules of biological interest by way of 3D-EM also passes review.

For the example of mixtures of known distributions as well as for positron emission tomography, there is a well-rounded theory for the convergence of the EM algorithm to wit the alternating projection approach of Csiszár and Tusnády [23] and the majorizing functional approach of Mülthei and Schorr [75] and De Pierro [29]. This approach extends to EM-like algorithms for some maximum likelihood-like problems. Unfortunately, this ignores the fact that the maximum likelihood problem is ill conditioned when the number of components in the mixture is large (or that the deconvolution problem is ill-posed). So, one needs to regularize the maximum likelihood problems, and then, in this chapter, the issue is whether there are EM algorithms for the regularized problems. For the PET problem, this certainly works for Bayesian approaches, leading to maximum a posteriori (MAP) likelihood problems as well as to arbitrary convex maximum

penalized likelihood problems. In this context, mention should be made of the EMS algorithm of Silverman et al. [90], the EM algorithm with a linear smoothing step added, and the NEMS algorithm of Eggermont and LaRiccia [36] in which an extra nonlinear smoothing step is added to the EMS algorithm to make it a genuine EM algorithm for a smoothed maximum likelihood problem. However, the convergence of regularization procedures for ill-posed maximum likelihood estimation problems, whether Tikhonov style penalization or "optimally" stopping the EM algorithm, will not be discussed. See, e.g., [80].

The final issue under consideration is that EM algorithms are painfully slow, so methods for accelerating EM algorithms are discussed as well. The accelerated methods take the form of block-iterative methods, including the extreme case of row-action methods.

The selection of topics is driven by applications to image processing. As such, there is very little overlap with the extensive up-to-date survey of EM algorithms of McLachlan and Krishnan [71].

2 The Kullback–Leibler Divergence

Before turning to the issue of EM algorithms, emphatic mention must be made of the pervasiveness of the Kullback–Leibler divergence (also called I-divergence or information divergence; see, e.g., [22] and references therein) in maximum likelihood estimation. For probability density functions f and g on $\mathbb{R}^d$, say, it is defined as

$$\mathrm{KL}(f,g) = \int_{\mathbb{R}^d} \left(f(y) \log \left\{ \frac{f(y)}{g(y)} \right\} + g(y) - f(y) \right) d\mu(y), \tag{10}$$

with μ denoting Lebesgue measure. Here, $0 \log(0/0)$ is defined as 0. Note that the Kullback–Leibler divergence is not symmetric in its arguments. Also note that the integrand is nonnegative, so that the integral is well defined if the value $+\infty$ is admitted. Moreover, the integrand equals 0 if and only if $f(y) = g(y)$, so that $\mathrm{KL}(f,g) > 0$ unless $f = g$ almost everywhere, in which case $\mathrm{KL}(f,g) = 0$.

Now consider the problem of estimating the unknown parameter x_o in a probability density $f(\,\cdot\,|x_o)$. In view of the above, the ideal way would be to

$$\text{minimize} \quad \mathrm{KL}\big(f(\,\cdot\,|x_o), f(\,\cdot\,|x)\big) \quad \text{over} \quad x, \tag{11}$$

but of course, this is not a rational problem because the objective function is unknown. However, note that

$$\begin{aligned} \mathrm{KL}\big(f(\,\cdot\,|x_o), f(\,\cdot\,|x)\big) = &- \int_{\mathbb{R}^d} f(y|x_o) \log f(y|x)\, d\mu(y) \\ &+ \int_{\mathbb{R}^d} f(y|x_o) \log f(y|x_o)\, d\mu(y), \end{aligned} \tag{12}$$

and that the second term does not depend on x. So, the problem (11) is equivalent to (has the same solutions as)

$$\text{minimize} \quad -\int_{\mathbb{R}^d} f(y|x_o)\, \log f(y|x)\, d\mu(y) \quad \text{over} \quad x. \tag{13}$$

Of course, this is still not a rational problem, but the objective function equals $\mathbb{E}[L_n(x)]$, where $L_n(x)$ is the scaled negative log-likelihood

$$L_n(x) = -\frac{1}{n} \sum_{i=1}^{n} \log f(Y_i|x), \tag{14}$$

if $Y_1, Y_2, \ldots, Y_n$ is a random sample of the random variable Y with probability density function $f(\,\cdot\,|x_o)$. So, solving the maximum likelihood problem (1) may be viewed as approximately solving the minimum Kullback–Leibler divergence problem (11). This is the basic reason for the pervasiveness of the Kullback–Leibler divergence in the analysis of maximum likelihood estimation and EM algorithms.

The above illustrates two additional points. First, in maximum likelihood estimation, one attempts to solve the minimum Kullback–Leibler problem (11) by first estimating the objective function. So, if the estimator is "optimal" at all, it has to be in a sense related to the Kullback–Leibler divergence. Second, one may well argue that one is not estimating the parameter x_o but rather the density $f(\,\cdot\,|x_o)$. This becomes especially clear if $f(\,\cdot\,|x)$ is reparametrized as $\varphi(\,\cdot\,|z) = f\big(\,\cdot\,|T(z)\big)$ for some transformation T. This would have an effect on the possible unbiasedness of the estimators $\hat{x}$ and $\hat{z}$ of x_o and z_o. However, under reasonable conditions on T, the maximum likelihood density estimators $f(\,\cdot\,|\,\hat{x})$ and $\varphi(\,\cdot\,|\,\hat{z})$ of $f(\,\cdot\,|\,x_o)$ will be the same.

3 The EM Algorithm

The Maximum Likelihood Problem

Let $(\mathcal{Y}, \mathcal{B}_\mathcal{Y}, \mathcal{P})$ be a statistical space, i.e., $(\mathcal{Y}, \mathcal{B}_\mathcal{Y})$ is a measurable space and $\mathcal{P}$ is a collection of probability measures on $\mathcal{B}_\mathcal{Y}$, represented as a family indexed by some index set $\mathcal{X}$ as follows:

$$\mathcal{P} = \{P(\,\cdot\,|x) : x \in \mathcal{X}\}. \tag{15}$$

Assume that there is a measure P_∞ that dominates $\mathcal{P}$ in the sense that every $P(\,\cdot\,|x)$ is absolutely continuous with respect to P_∞. Then, the Radon–Nikodym derivative of $P(\,\cdot\,|x)$ with respect to P_∞ exists and is $\mathcal{B}_\mathcal{Y}$ measurable for all $x \in \mathcal{X}$. It may be written as

$$f_\mathcal{Y}(y|x) = \left[\frac{dP(\,\cdot\,|x)}{dP_\infty}\right](y) \tag{16}$$

and is referred to as the density of $P(\,\cdot\,|x)$ with respect to P_∞. It should be observed that $f_{\mathcal{Y}}(\,\cdot\,|x_o) = f_Y(\,\cdot\,)$ is the density of the random variable Y with respect to P_∞. For arbitrary x, $f_{\mathcal{Y}}(\,\cdot\,|x)$ is a density, but since it is not known of what random variable, the subscript $\mathcal{Y}$ is used here.

Let Y be a random variable with values in $\mathcal{Y}$, and assume that it is distributed as $P(\cdot\,|x_o)$ for some (unknown) $x_o \in \mathcal{X}$. The objective is to estimate x_o based on a random sample $Y_1, Y_2, \ldots, Y_n$ of the random variable Y. Note that estimating x_o amounts to constructing a measurable function of the data, which may then be denoted as $\hat{x} = \hat{x}(Y_1, Y_2, \ldots, Y_n)$.

The maximum likelihood problem for estimating x_o is then written in the form of minimizing the scaled negative log-likelihood,

$$\begin{aligned} &\text{minimize} \quad L_n(x) \stackrel{\text{def}}{=} -\frac{1}{n}\sum_{i=1}^{n} \log f_{\mathcal{Y}}(Y_i|x) \\ &\text{subject to} \quad x \in \mathcal{X}. \end{aligned} \tag{17}$$

In this formulation, the parameter x is deemed important. The alternative formulation in which the densities are deemed important is

$$\begin{aligned} &\text{minimize} \quad \tilde{L}_n(f) \stackrel{\text{def}}{=} -\frac{1}{n}\sum_{i=1}^{n} \log f(Y_i) \\ &\text{subject to} \quad f \in \mathcal{P}. \end{aligned} \tag{18}$$

In this formulation, there are two ingredients: the likelihood function and the (parametric) family of densities under consideration.

It is not obvious that solutions should exist, especially if the index set $\mathcal{X}$ is large, but in applications of image processing type, this turns out to be of lesser importance than one might think. See Sect. 7. Regardless, closed form solutions are generally not available, and one must employ iterative methods for the solution of the maximum likelihood problem. In this chapter, that means EM algorithms.

The Bare-Bones EM Algorithm

Here, the bare essentials of the EM algorithm are presented. The basic premise in the derivation of the EM algorithm is that there is "missing" data that would make estimating x_o a lot easier had they been observed. So, assume that the missing data refers to data in a space $\mathcal{Z}$, with $(\mathcal{Z}, \mathcal{B}_{\mathcal{Z}}, \mathcal{Q})$ another statistical space, where the collection of probability measures is again indexed by $\mathcal{X}$,

$$\mathcal{Q} = \{Q(\,\cdot\,|x) : x \in \mathcal{X}\}. \tag{19}$$

Assume that $\mathcal{Q}$ is dominated by some measure Q_∞, and denote the associated Radon–Nikodym derivatives as

$$f_{\mathcal{Z}}(z|x) = \left[\frac{dQ(\,\cdot\,|x)}{dQ_\infty}\right](z), \quad z \in \mathcal{Z}. \tag{20}$$

Let Z be a random variable with values in $\mathcal{Z}$ and with distribution $Q(\,\cdot\,|x_o)$, with the same x_o as for the random variable Y. Note that the pair (Y, Z) takes on values in $\mathcal{Y} \times \mathcal{Z}$. The relevant statistical space is then $(\mathcal{Y} \times \mathcal{Z}, B_{\mathcal{Y}\times Z}, \mathcal{R})$, where $B_{\mathcal{Y}\times\mathcal{Z}}$ is the smallest σ-algebra that contains all sets $A \times B$ with $A \in \mathcal{B}_{\mathcal{Y}}$ and $B \in \mathcal{B}_{\mathcal{Z}}$. Again, assume that $\mathcal{R}$ may be indexed by $\mathcal{X}$ as

$$\mathcal{R} = \{R(\,\cdot\,|x) : x \in \mathcal{X}\}, \tag{21}$$

and that $\mathcal{R}$ is dominated by some measure R_∞. Write

$$f_{\mathcal{Y},\mathcal{Z}}(y, z|x) = \left[\frac{dR(\,\cdot\,|x)}{dR_\infty}\right](y, z) \tag{22}$$

for the associated Radon–Nikodym derivatives.

Now, if the "complete" data $(Y_1, Z_1), (Y_2, Z_2), \ldots, (Y_n, Z_n)$, a random sample of the random variable (Y, Z), is available, then the maximum likelihood problem for estimating x_o is

$$\begin{aligned} &\text{minimize} \quad -\frac{1}{n}\sum_{i=1}^{n} \log f_{\mathcal{Y},\mathcal{Z}}(Y_i, Z_i|x) \\ &\text{subject to} \quad x \in \mathcal{X}. \end{aligned} \tag{23}$$

Of course, this is not a rational problem, since the Z_i went unobserved. In other words, the objective function is not known (and not knowable). However, one may attempt to estimate it by the conditional expectation

$$\mathbb{E}\left[-\frac{1}{n}\sum_{i=1}^{n} \log f_{\mathcal{Y},\mathcal{Z}}(Y_i, Z_i|x) \,\middle|\, \mathbb{Y}_n\right] = -\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\log f_{\mathcal{Y},\mathcal{Z}}(Y_i, Z_i|x) \,\middle|\, Y_i\right],$$

where $\mathbb{Y}_n = (Y_1, Y_2, \ldots, Y_n)$. The fly in the ointment is that computing this expectation involves the distribution of Z conditioned on Y, which surely will involve the unknown x_o one wishes to estimate! So, at this point, assume that some initial guess x_1 for x_o is available; then denote the resulting (approximate) conditional expectation by $\mathbb{E}[\ldots|\mathbb{Y}_n, x_1]$.

Determining this conditional expectation constitutes the E-step of the first iteration of the EM algorithm. The M-step of the first iteration then amounts to solving the minimization problem

$$\begin{aligned} &\text{minimize} \quad \Lambda_n(x|x_1) \stackrel{\text{def}}{=} -\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\log f_{\mathcal{Y},\mathcal{Z}}(Y_i, Z_i|x)\,|x_1, Y_i] \\ &\text{subject to} \quad x \in \mathcal{X}. \end{aligned} \tag{24}$$

Denote the solution by x_2 (assuming it exists). Suppressing the presence of the Y_i in the notation, one may define the iteration operator by $x_2 = R(x_1)$, and then the EM algorithm may be stated as

$$x_{k+1} = R(x_k), \quad k = 1, 2, \ldots, \tag{25}$$

provided x_1 has been chosen appropriately. This is the bare-bones version of the EM algorithm. Note that it may not be necessary to solve the problem (24) exactly, e.g., one may consider (over)relaxation ideas or the so-called stochastic EM algorithms. See, e.g., [76] and references therein. This will not be considered further.

Remark 1. (a) It may be inappropriate to speak of the EM algorithm, since the introduction of different missing data may lead to a different algorithm. However, usually there is not much choice in the missing data.

(b) There is a different approach to the complete data, by assuming that $Y = T(Z)$ for some many-to-one map $T : \mathcal{Z} \to \mathcal{Y}$. Then, Z is the complete data and Y is the incomplete data, but one does not identify missing data as such.

The Bare-Bones EM Algorithm Fleshed Out

Here, some of the details of the bare-bones EM algorithm are filled in by using explicit expressions for the conditional expectations. To that end, assume that one may take the dominating measure R_∞ to be $R_\infty = P_\infty \cdot Q_\infty$, in the sense that

$$R_\infty(A \times B) = P_\infty(A) \cdot Q_\infty(B) \quad \text{for all } A \in \mathcal{B}_\mathcal{Y} \text{ and } B \in \mathcal{B}_\mathcal{Z}. \tag{26}$$

Let (Y, Z) have density $f_{\mathcal{Y},\mathcal{Z}}(y, z|x_o)$ with respect to the product measure $P_\infty \times Q_\infty$. Then, for all $\mathcal{A} \in \mathcal{B}_{\mathcal{Y}\times\mathcal{Z}}$ and all measurable functions h on $\mathcal{Y} \times \mathcal{Z}$ with finite expectation $\mathbb{E}[h(Y, Z)]$, one may write

$$\begin{aligned} \mathbb{E}[h(Y,Z)] &= \int_{\mathcal{A}} h(y,z)\, f_{\mathcal{Y},\mathcal{Z}}(y,z|x)\, dP_\infty(y)\, dQ_\infty(z) \\ &= \int_{\mathcal{Y}} \left\{ \int_{\mathcal{Z}} h(y,z)\, f_{\mathcal{Y},\mathcal{Z}}(y,z|x)\, dQ_\infty(z) \right\} dP_\infty(y) \qquad \text{(Fubini)} \\ &= \int_{\mathcal{Y}} \left\{ \int_{\mathcal{Z}} h(y,z)\, \frac{f_{\mathcal{Y},\mathcal{Z}}(y,z|x)}{f_\mathcal{Y}(y|x)}\, dQ_\infty(z) \right\} f_\mathcal{Y}(y|x)\, dP_\infty(y). \end{aligned}$$

It is clear that this may be interpreted as the expected value of

$$\int_{\mathcal{Z}} h(Y,z)\,\frac{f_{\mathcal{Y},\mathcal{Z}}(Y,z|x)}{f_{\mathcal{Y}}(Y|x)}\,dQ_\infty(z),$$

and then interpret this in turn as $\mathbb{E}[h(Y,Z)\,|\,Y]$, the expected value of $h(Y,Z)$ conditioned on Y.

Now define the density of Z conditional on Y by

$$f_{\mathcal{Z}|\mathcal{Y}}(z|y,x) = \frac{f_{\mathcal{Y},\mathcal{Z}}(y,z|x)}{f_{\mathcal{Y}}(y|x)} \tag{27}$$

for those y for which $f_{\mathcal{Y}}(y|x) > 0$ (and arbitrarily if $f_{\mathcal{Y}}(y|x) = 0$). Similarly, one defines

$$f_{\mathcal{Y}|\mathcal{Z}}(y|z,x) = \frac{f_{\mathcal{Y},\mathcal{Z}}(y,z|x)}{f_{\mathcal{Z}}(z|x)} \tag{28}$$

for those z for which $f_{\mathcal{Z}}(z|x) > 0$ (and arbitrarily if $f_{\mathcal{Z}}(z|x) = 0$). So then Bayes' rule yields

$$f_{\mathcal{Z}|\mathcal{Y}}(z|y,x) = \frac{f_{\mathcal{Y}|\mathcal{Z}}(y|z,x)\,f_{\mathcal{Z}}(z|x)}{f_{\mathcal{Y}}(y|x)}. \tag{29}$$

The conditional expectation of a measurable function $h(Y,Z)$ given Y is then

$$\mathbb{E}[h(Y,Z)|Y,x] = \int_{\mathcal{Z}} h(Y,z)\,f_{\mathcal{Z}|\mathcal{Y}}(z|Y,x)\,dQ_\infty(z). \tag{30}$$

Probabilists force us to add "almost surely" here.

Now apply this to the conditional expectation of $\log f_{\mathcal{Y},\mathcal{Z}}(Y,Z|x)$, with a guess x_1 of the true x. Then,

$$\begin{aligned}
&\mathbb{E}[\log f_{\mathcal{Y},\mathcal{Z}}(Y,Z|x)|Y,x_1] \\
&\quad= \mathbb{E}[\log f_{\mathcal{Z}}(Z|x) + \log f_{\mathcal{Y}|\mathcal{Z}}(Y|Z,x)|Y,x_1] \\
&\quad= \int_{\mathcal{Z}} \frac{f_{\mathcal{Y}|\mathcal{Z}}(Y|z,x_1)\,f_{\mathcal{Z}}(z|x_1)}{f_{\mathcal{Y}}(y|x_1)}\,\log f_{\mathcal{Z}}(z|x)\,dQ_\infty(z) \\
&\qquad+ \int_{\mathcal{Z}} \frac{f_{\mathcal{Y}|\mathcal{Z}}(Y|z,x_1)\,f_{\mathcal{Z}}(z|x_1)}{f_{\mathcal{Y}}(Y_i|x_1)}\,\log f_{\mathcal{Y}|\mathcal{Z}}(Y|z,x)\,dQ_\infty(z).
\end{aligned} \tag{31}$$

For $\Lambda_n(x|x_1)$, this gives

$$\Lambda_n(x|x_1) = \int_{\mathcal{Z}} \varphi_{\mathcal{Z}}(z|x_1)\, \log f_{\mathcal{Z}}(z|x)\, dQ_\infty(z) + \\ -\frac{1}{n} \sum_{i=1}^{n} \int_{\mathcal{Z}} \frac{f_{\mathcal{Y}|\mathcal{Z}}(Y_i|z, x_1)\, f_{\mathcal{Z}}(z|x_1)}{f_{\mathcal{Y}}(Y_i|x_1)}\, \log f_{\mathcal{Y}|\mathcal{Z}}(Y_i|z, x)\, dQ_\infty(z), \quad (32)$$

with

$$\varphi_{\mathcal{Z}}(z) = f_{\mathcal{Z}}(z|x_1) \cdot \frac{1}{n} \sum_{i=1}^{n} \frac{f_{\mathcal{Y}|\mathcal{Z}}(Y_i|z, x_1)}{f_{\mathcal{Y}}(Y_i|x_1)}. \quad (33)$$

For the M-step of the algorithm, one has to minimize this over x, which is in general not trivial. The problem is simplified somewhat in the important case where $f_{\mathcal{Y}|\mathcal{Z}}(y|z, x)$ is known and does not depend on x. Then, the problem reduces to solving

$$\begin{aligned} &\text{minimize} \quad \mathfrak{L}_n(x|x_1) \stackrel{\text{def}}{=} -\int_{\mathcal{Z}} \varphi_{\mathcal{Z}}(z|x_1)\, \log f_{\mathcal{Z}}(z|x)\, dQ_\infty(z) \\ &\text{subject to} \quad x \in \mathcal{X}. \end{aligned} \quad (34)$$

Note that

$$\mathfrak{L}_n(x|x_1) = \mathrm{KL}\left(\varphi_{\mathcal{Z}}(\,\cdot\,|x)\,,\ f_{\mathcal{Z}}(\,\cdot\,, x)\right) + (\text{terms not depending on } x), \quad (35)$$

where $\mathrm{KL}(f, g)$ is the Kullback–Leibler divergence between the density f and g with respect to the same measure Q_∞, defined as

$$\mathrm{KL}(f, g) = \int_{\mathcal{Z}} \left\{ f(z) \log \left\{ \frac{f(z)}{g(z)} \right\} + f(z) - g(z) \right\} dQ_\infty(z). \quad (36)$$

Compare with (10).

So, solving (34) amounts to computing what one may call the Kullback–Leibler projection of $\varphi_{\mathcal{Z}}(\,\cdot\,|x_1)$ onto the parametric family $\mathcal{P}$. If $\mathcal{P}$ is such that $\varphi_{\mathcal{Z}}(\,\cdot\,|x_1) \in \mathcal{P}$, then the projection is $f_{\mathcal{Z}}(\,\cdot\,|x) = \varphi_{\mathcal{Z}}(\,\cdot\,|x_1)$.

The EM Algorithm Increases the Likelihood

The expression for $\Lambda_n(x|x_1)$ (see (24)) may be reworked as follows. Using

$$f_{\mathcal{Y},\mathcal{Z}}(y, z, |x) = f_{\mathcal{Z}|\mathcal{Y}}(z|y, x)\, f_{\mathcal{Y}}(y|x)$$

(see (27)), one gets

$$\mathbb{E}\left[\log f_{\mathcal{Y},\mathcal{Z}}(Y,Z|x) \;\middle|\; x_1, Y\right] = \log f_{\mathcal{Y}}(Y|x) + \mathbb{E}[\log f_{\mathcal{Z}|\mathcal{Y}}(Z|Y,x)\,|Y,x_1],$$

and so,

$$\Lambda(x|x_1) = L_n(x) + e_n(x_1|x), \tag{37}$$

where $L_n(x)$ is given by (17) and

$$e_n(u|w) \stackrel{\text{def}}{=} -\sum_{i=1}^{n} \int_{\mathcal{Z}} f_{\mathcal{Z}|\mathcal{Y}}(z|Y_i,u)\, \log f_{\mathcal{Z}|\mathcal{Y}}(z|Y_i,w)\, dQ_\infty(z). \tag{38}$$

It is now obvious that the EM algorithm decreases $L_n(x)$: Let x_2 be the minimizer of $\Lambda(x|x_1)$ over $x \in \mathcal{X}$. Then, $\Lambda_n(x_2|x_1) \leqslant \Lambda_n(x_1|x_1)$, and so $L_n(x_2) + e_n(x_1|x_2) \leqslant L_n(x_1) + e_n(x_1|x_1)$, or

$$L_n(x_1) - L_n(x_2) \geqslant e_n(x_1|x_2) - e_n(x_1|x_1) = K_n(x_1|x_2), \tag{39}$$

where

$$K_n(u|w) = \sum_{i=1}^{n} \mathrm{KL}\left(f_{\mathcal{Z}|\mathcal{Y}}(\,\cdot\,|Y_i,u), f_{\mathcal{Z}|\mathcal{Y}}(\,\cdot\,|Y_i,w)\right) \tag{40}$$

is a sum of Kullback–Leibler "distances"; see (36). Then, $K_n(x_1|x_2) \geqslant 0$ unless $x_1 = x_2$, assuming that the conditional densities $f_{Z|Y}(\,\cdot\,|Y,u)$ and $f_{Z|Y}(\,\cdot\,|Y,w)$ are equal Q_∞ almost everywhere only if $u = w$. Then, the conclusion

$$L_n(x_1) > L_n(x_2) \quad \text{unless} \quad x_1 = x_2 \tag{41}$$

is justified. Thus, the EM algorithm decreases the likelihood.

Unfortunately, $x_1 = x_2$ being a fixed point of the EM iteration does not guarantee that then x_1 is a maximum likelihood estimator of x_o. Equally unfortunately, even if one gets an infinite sequence of estimators, this does not imply that the sequence of estimators converges, nor that the likelihood converges to its maximum. Later on, the convergence of EM algorithms for special, convex maximum likelihood problems is discussed in detail.

4 The EM Algorithm in Simple Cases

In this section, some simple cases of the EM algorithm are discussed, capturing some of the essential features of more complicated "real" examples of maximum likelihood estimation to be discussed later on. It turns out that the EM algorithms are the "same" in all but the last example (regarding a finite mixture of unknown

densities), even though the settings appear to be quite different. However, even in the last case, the "same" EM algorithm arises via the empirical Bayes approach.

A word on notation: The scaled negative log-likelihood for each problem is always denoted as L_n; $L_n(x)$ in the discrete case, $L_n(f)$ in the continuous case. The negative log-likelihood in the M-step of the EM algorithm is denoted by $\Lambda(x|x_1)$ or variations thereof.

Mixtures of Known Densities

Let $d \geqslant 1$ be an integer and consider the statistical space $(\mathcal{Y}, \mathcal{B}, \mathcal{P})$, with $\mathcal{Y} = \mathbb{R}^d$, $\mathcal{B}$ the σ-algebra of Borel subsets of $\mathcal{Y}$ and $\mathcal{P}$ the collection of probability measures that are absolutely continuous with respect to Lebesgue measure. Consider the random variable Y with values in $\mathcal{Y}$ and with density

$$f_Y(y) = \sum_{j=1}^{m} x_o(j)\, a_j(y), \quad y \in \mathcal{Y}, \tag{42}$$

where $a_1, a_2, \ldots, a_m$ are known densities and $x_o = (x_{o,1}, x_{o,2}, \ldots, x_{o,m})^{\mathrm{T}}$ is a probability vector, i.e., $x_o \in V_m$,

$$V_m = \left\{ x \in \mathbb{R}^m \,\middle|\, x \geqslant 0 \text{ (componentwise)}, \sum_{j=1}^{m} x(j) = 1 \right\}. \tag{43}$$

Suppose that one has a random sample $Y_1, Y_2, \ldots, Y_n$ of the random variable Y. The maximum likelihood problem for estimating f_Y (or estimating the probability vector x_o) is then

$$\begin{aligned} &\text{minimize} \quad L_n(x) \stackrel{\text{def}}{=} -\frac{1}{n} \sum_{i=1}^{n} \log \left(\sum_{j=1}^{m} x(j)\, a_j(Y_i) \right) \\ &\text{subject to} \quad x \in V_m. \end{aligned} \tag{44}$$

To derive an EM algorithm, missing data must be introduced. To see what could be missing, it is helpful to think of how one would simulate the random variable Y. First, draw the random variable J from the distribution

$$f_J(j) = \mathbb{P}[J = j] = x_o(j), \quad j = 1, 2, \ldots, m. \tag{45}$$

Then, conditional on $J = j$, draw Y from the distribution with density a_j. So, the missing data is J, a random variable with values in $\mathcal{M} = \{1, 2, \ldots, m\}$. The associated statistical space is

$$\left(\mathcal{M}, 2^M, V_m\right), \tag{46}$$

the σ-algebra is the collection of all subsets of $\mathcal{M}$, and the collection of all probability measures on $\mathcal{M}$ may be represented by V_m. Let α denote counting measure on $\mathcal{M}$, i.e., for any $A \in 2^{\mathcal{M}}$,

$$\alpha(A) = |\,A\,| \quad \text{(the number of elements in } A\text{)}. \tag{47}$$

Then, it is easy to see that the distribution of (Y, J) is absolutely continuous with respect to the product measure $\mu \times \alpha$, with density

$$f_{Y,J}(y, j) = f_J(j)\, f_{Y|J}(y|j) = x_o(j)\, a_j(y), \quad y \in \mathcal{Y},\ j \in \mathcal{M}. \tag{48}$$

Now, the complete data is $(Y_1, J_1), (Y_2, J_2), \ldots, (Y_n, J_n)$ and the complete maximum likelihood problem for estimating x_o is

$$\text{minimize} \quad -\frac{1}{n} \sum_{i=1}^{n} \log \left\{ x(J_i)\, a_{J_i}(Y_i) \right\} \quad \text{subject to} \quad x \in V_m. \tag{49}$$

Of course, the J_i went unobserved, so one must compute the conditional expectations $\mathbb{E}\left[\log \left\{x(J)\, a_J(Y)\right\} \,\middle|\, Y\right]$. Now,

$$f_{J|Y}(j|y) = \frac{f_{Y,J}(y, j)}{f_Y(y)} = \frac{x_o(j)\, a_j(y)}{\sum_{p=1}^{m} a_p(y)\, x_o(j)},$$

but of course, x_o is unknown; approximate it by some initial guess $x^{[1]} \in V_m$, e.g., $x_j^{[1]} = 1/m$ for all j. Then, the conditional expectation in question is approximated by

$$\begin{aligned} \mathbb{E}\left[\log \left\{x(J)\, a_J(Y)\right\} \,\middle|\, Y,\, x^{[1]}\right] &= -\int_{\mathcal{M}} \log \left\{x(j)\, a_j(Y)\right\}\, x^{[2]}(j, Y)\, d\alpha(j) \\ &= -\sum_{j \in \mathcal{M}} x^{[2]}(j, Y)\, \log \left\{x(j)\, a_j(Y)\right\}, \end{aligned}$$

with $$x^{[2]}(j, Y) = \frac{x^{[1]}(j)\, a_j(Y)}{\sum_{p=1}^{m} a_p(Y)\, x^{[1]}(p)}, \quad j \in \mathcal{M}.$$

Then, the E-step of the EM algorithm leads to the problem

$$\text{minimize} \quad -\sum_{j \in \mathcal{M}} x^{[2]}(j)\, \log \left\{x(j)\, a_{ij}\right\} \quad \text{subject to} \quad x \in V_\ell, \tag{50}$$

where $a_{ij} = a_j(Y_i)$ for all i, j and $x^{[2]}(j) = \frac{1}{n}\sum_{i=1}^{n} x^{[2]}(j, Y_i)$, or

$$x^{[2]}(j) = x^{[1]}(j) \cdot \frac{1}{n} \sum_{i=1}^{n} \frac{a_{ij}}{\left(\sum_{p=1}^{m} a_{ip}\, x^{[1]}(p) \right)}. \tag{51}$$

Taking into account that

$$\log \left\{ x(j)\, a_j(Y) \right\} = \log x(j) + (\text{a term not depending on } x),$$

one then arrives at the problem

$$\begin{aligned} &\text{minimize} \quad \Lambda_n \left(x | x^{[2]} \right) \stackrel{\text{def}}{=} -\sum_{j=1}^{m} x^{[2]}(j) \log x(j) \\ &\text{subject to} \quad x \in V_m. \end{aligned} \tag{52}$$

This is the E-step of the first iteration of the EM algorithm. Now consider the identity

$$\Lambda \left(x | x^{[2]} \right) - \Lambda \left(x^{[2]} | x^{[2]} \right) = \mathrm{KL} \left(x^{[2]}, x \right), \tag{53}$$

where for u, w nonnegative vectors in $\mathbb{R}^m$,

$$\mathrm{KL}(u, w) \stackrel{\text{def}}{=} \sum_{j=1}^{m} \left\{ u(j) \, \log \frac{u(j)}{w(j)} + w(j) - u(j) \right\}. \tag{54}$$

This is the finite-dimensional Kullback–Leibler divergence between the nonnegative vectors u and w. Note that the summand is nonnegative and so is minimal when $u = w$. So, the solution of (52) is precisely $x = x^{[2]}$. This would be the M-step of the first iteration of the EM algorithm. The EM-step is then (51).

A Deconvolution Problem

The setting is the statistical space $\left(\mathbb{R}^d, \mathcal{B}, \mathcal{P} \right)$, where $\mathcal{B}$ is the σ-algebra of Borel subsets of $\mathbb{R}^d$ and $\mathcal{P}$ is the collection of all probability density functions on $\mathbb{R}^d$ (with respect to Lebesgue measure). Denote Lebesgue measure by μ. Let Y be a random variable with values in $\mathbb{R}^d$ and with density f_Y (with respect to Lebesgue measure); the interest is in estimating f_Y. Now, assume that one is unable to observe Y directly but that instead one only observes a corrupted version, viz., $W = Y + Z$, where Z is another $\mathbb{R}^d$-valued random variable. Assume that the distribution of Z is completely known; denote its density by k. The density of W is then $\mathcal{K} f_Y$, where the integral operator $\mathcal{K} : L^1 \left(\mathbb{R}^d, d\mu \right) \to L^1 \left(\mathbb{R}^d, d\mu \right)$ is defined as

$$[\mathcal{K} f](w) = \int_{\mathbb{R}^d} k(w-y)\, f(y)\, d\mu(y), \quad w \in \mathbb{R}^d. \tag{55}$$

Note that $k(\cdot - y)$ is the density of W conditioned on $Y = y$, i.e., $f_{W|Y}(w|y) = k(w-y)$ for all w and y, and then

$$f_{W,Y}(w, y) = k(w-y)\, f_Y(y)\,, \quad w,\ y \in \mathbb{R}^d. \tag{56}$$

So, assume that one has a random sample $W_1, W_2, \ldots, W_n$ of the random variable W. The maximum likelihood problem for estimating f_Y is then

$$\begin{aligned} &\text{minimize} \quad L_n(f) \stackrel{\text{def}}{=} -\frac{1}{n} \sum_{i=1}^{n} \log[\mathcal{K} f](W_i) \\ &\text{subject to} \quad f \in \mathcal{P}. \end{aligned} \tag{57}$$

Recall that $\mathcal{P}$ is the collection of all pdfs in $L^1(\mathbb{R}^d, d\mu)$. It is impossible to guarantee that this problem has a solution. In fact, regularization is required; see section "Smoothed EM Algorithms." Nevertheless, one can use EM algorithms to construct useful approximations to the density f_Y by the expedient of stopping the iteration "early."

Note that one could view (57) as a continuous mixture problem, since $\mathcal{K} f$ is a continuous mixture of known densities to wit the known densities $k_y(w) = k(w-y)$, $y \in \mathbb{R}^d$, and the continuous weights are the unknown $f_Y(y)$, $y \in \mathbb{R}^d$. However, the present approach is somewhat different.

To derive an EM algorithm, one must decide on the missing data. It seems obvious that the missing data is Y itself or Z (or both), but the choice Y seems the most convenient. Thus, assume that one has available the random sample $(W_1, Y_1), (W_2, Y_2), \ldots, (W_n, Y_n)$ of the random variable (W, Y). In view of (56), the maximum likelihood problem for estimating f_Y is then

$$\begin{aligned} &\text{minimize} \quad \Lambda_n(f) \stackrel{\text{def}}{=} -\frac{1}{n} \sum_{i=1}^{n} \log\{k(W_i - Y_i)\, f(W_i)\} \\ &\text{subject to} \quad f \in \mathcal{P}. \end{aligned} \tag{58}$$

Since the Y_i are not really observed, one must compute or approximate the conditional expectation $\mathbb{E}\left[\log\{k(W-Y)\, f(Y)\} \,\middle|\, W\right]$. Since the density of Y conditioned on W may be written as

$$f_{Y|W}(y|w) = \frac{k(w-y)\, f_Y(y)}{[\mathcal{K} f_Y](w)},$$

then, approximating f_Y by some initial guess f_1, the conditional expectation is approximated by

$$\int_{\mathbb{R}^d} \frac{k(W-y)\, f_1(y)}{[\mathcal{K} f_1](w)} \log f(y)\, d\mu(y),$$

apart from a term not depending on f. So then, the problem (58) is approximated by

$$\text{minimize} \quad -\int_{\mathbb{R}^d} f_2(y) \log f(y)\, d\mu(y) \quad \text{subject to} \quad f \in \mathcal{P}, \tag{59}$$

where

$$f_2(y) = f_1(y) \cdot \frac{1}{n} \sum_{i=1}^{n} \frac{k(W_i - y)}{[\mathcal{K} f_1](W_i)}, \quad y \in \mathbb{R}^d. \tag{60}$$

This is the E-step of the EM algorithm

For the M-step, i.e., actually solving (59), note that

$$\Lambda_n(f \,|\, f_1) - \Lambda_n(f_1 \,|\, f_1) = \mathrm{KL}(f_2, f), \tag{61}$$

which is minimal for f_2. Thus, the solution of (59) is $f = f_2$ as well. Thus, the EM algorithm takes the form (60) iteratively applied.

The discretized EM algorithm: The EM algorithm (60) cannot be implemented as is, but it certainly may be discretized. However, it is more straightforward to discretize the maximum likelihood problem (57).

A reasonable way to discretize the maximum likelihood problem (57) is to restrict the minimization to step functions on a suitable partition of the space. Suppose that the compact set $C_o \subset \mathbb{R}^d$ contains the support of f_Y, and let $\{C_j\}_{j=1}^m$ be a partition of C_o. Define the (step) functions a_j by

$$a_j(y) = |\, C_j \,|^{-1}\, \mathbf{1}(y \in C_j)\,, \quad i = 1, 2, \cdot, m, \tag{62}$$

where for any set A, the indicator function $\mathbf{1}(y \in A)$ is defined as

$$\mathbf{1}(y \in A) = 1 \quad \text{if} \quad y \in A \quad \text{and} \quad = 0 \quad \text{otherwise.} \tag{63}$$

Then, define $\mathcal{P}_m$ to be the set of pdfs in the linear span of the a_j,

$$\mathcal{P}_m = \left\{ \sum_{j=1}^{m} x_j\, a_j(\,\cdot\,) \,\middle|\, x \in V_m \right\}. \tag{64}$$

Note that the a_j are pdfs, and in fact, one could take the a_j, $j = 1, 2, \ldots, m$ to be *any* collection of pdfs.

Now, consider the restricted maximum likelihood problem

$$\text{minimize} \quad -\frac{1}{n}\sum_{i=1}^{n} \log[\mathcal{K}f](W_i) \quad \text{subject to} \quad f \in \mathcal{P}_m \tag{65}$$

and observe that it may obviously be rewritten as

$$\text{minimize} \quad -\frac{1}{n}\sum_{i=1}^{n} \log\left(\sum_{j=1}^{m} a_{ij}\, x_j\right) \quad \text{subject to} \quad x \in V_m, \tag{66}$$

where for $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, m$,

$$a_{ij} = \int_{\mathbb{R}^d} k(W_i - y)\, a_j(y)\, d\mu(y).$$

And this is all there is to it: The problem (66) is just a finite mixture problem with known distributions! Thus, the EM algorithm is as in section "Mixtures of Known Densities,"

$$x_j^{[k+1]} = x_j^{[k]} \cdot \frac{1}{n}\sum_{i=1}^{n} \frac{a_{ij}}{\left(\sum_{p=1}^{m} a_{ip}\, x_p^{[k]}\right)}, \quad j = 1, 2, \ldots, m, \tag{67}$$

with the estimator for f_Y induced by the representation of (64).

Another EM algorithm? There is of course another way to derive an EM algorithm for the problem (65), viz., by introducing the missing data Y_i as before. As for the unrestricted maximum likelihood problem (57), the E-step of the first iteration of the EM algorithm leads to the problem, analogous to (59),

$$\text{minimize} \quad -\int_{\mathbb{R}^d} f_2(y)\, \log f(y)\, d\mu(y) \quad \text{subject to} \quad f \in \mathcal{P}_m, \tag{68}$$

with f_2 given by (60). Now, using the representations

$$f(y) = \sum_{j=1}^{m} x_j\, a_j(y), \quad f_k(y) = \sum_{j=1}^{m} x_j^{[k]}\, a_j(y)$$

with the step functions a_j, for $k = 1, 2$, the objective function in (65) may be written as

$$-\sum_{j=1}^{m} (\log x_j) \int_{C_j} f_2(y)\, d\mu(y),$$

and of course

$$\int_{C_j} f_2(y)\, d\mu(y) = x_j^{[1]} \cdot \frac{1}{n} \sum_{i=1}^{n} \frac{a_{ij}}{[\mathcal{K} f_1](W_i)} \stackrel{\text{def}}{=} x_j^{[2]}, \tag{69}$$

where

$$a_{ij} = \int_{\mathbb{R}^d} k(W_i - y)\, a_j(y)\; d\mu(y) = |\, C_j \,|^{-1} \int_{C_j} k(W_i - y)\, d\mu(y).$$

Note that

$$[\mathcal{K} f_1](W_i) = \sum_{j=1}^{m} a_{ij}\, x_j^{[1]}\,, \quad i = 1, 2, \ldots, n.$$

Thus, the problem (68) is equivalent to

$$\text{minimize} \quad -\sum_{j=1}^{m} x_j^{[2]} \log x_j \quad \text{subject to} \quad x \geqslant 0\,,\ \sum_{j=1}^{m} x_j = 1. \tag{70}$$

But it was already shown in section "Mixtures of Known Densities" that the solution is $x = x^{[2]}$. So, the iterative step is *exactly* the same as in (67). As an aside, this is a case where the introduction of "different" missing data leads to the same EM algorithm.

The Deconvolution Problem with Binning

Consider again the deconvolution problem of section "A Deconvolution Problem," but now with the extra twist that the data is binned.

Recall that the random variable of interest is Y which lives in the statistical space $(\mathcal{Y}, \mathcal{B}_\mathcal{Y}, \mathcal{P})$ with $\mathcal{Y} = \mathbb{R}^d$, $\mathcal{B}_\mathcal{Y}$ the σ-algebra of Borel subsets of $\mathcal{Y}$, and $\mathcal{P}$ the collection of probability measures on $\mathcal{B}_\mathcal{Y}$ that are absolutely continuous with respect to Lebesgue measure. The density of Y is denoted by f_Y. The random variable Y was not observable. Instead, one can observe the random variable

$$W = Y + Z,$$

where Z is another random variable living in $(\mathcal{Y}, \mathcal{B}_\mathcal{Y}, \mathcal{P})$, with known density denoted by k and independent of Y. Actually, with binned data, W is not observed either. Let $\ell \in \mathbb{N}$ and let $\{B_j\}_{j=1}^{\ell} \subset \mathcal{B}_\mathcal{Y}$ be a partition of $\mathcal{Y}$ (or of a set containing the support of W). What one does observe is which "bin" B_j the observation W belongs to. That is, one observes the random variable J, with $J = j$ if

$$\mathbf{1}(W \in B_j) = 1, \tag{71}$$

cf. (63). Then, the statistical space of interest is $(\mathcal{M}, 2^{\mathcal{M}}, V_m)$; see (46). Of course, V_m is dominated by the counting measure, denoted by α; see (47). The density of J then satisfies

$$f_J(j) = [\mathcal{K} f_Y](B_j) = \int_{B_j} [\mathcal{K} f_Y](w)\, d\mu(w), \quad j \in \mathcal{M}. \tag{72}$$

So, now one observes the random sample $J_1, J_2, \ldots, J_n$ of the random variable J, and the goal is to estimate f_Y. The maximum likelihood problem is then

$$\text{minimize} \quad -\frac{1}{n} \sum_{i=1}^{n} \log[\mathcal{K} f](B_{J_i}) \quad \text{subject to} \quad f \in \mathcal{P}, \tag{73}$$

which is equivalent to

$$\text{minimize} \quad -\frac{1}{n} \sum_{j=1}^{\ell} N_j \, \log[\mathcal{K} f](B_j) \quad \text{subject to} \quad f \in \mathcal{P}. \tag{74}$$

Here, the N_j are the bin counts

$$N_j = \sum_{i=1}^{n} \mathbf{1}(J_i = j), \quad j \in \mathcal{M}. \tag{75}$$

Remark 2. Later, for any function $h : \mathcal{M} \to \mathbb{R}$, the more general identity

$$\sum_{i=1}^{n} h(J_i) = \sum_{j=1}^{m} N_j \, h(j)$$

will be useful.

So, the real data are the bin counts, but it is advantageous to keep the J_i. It has to be seen whether one can get away it, though. So, the starting point is (73) and not (74).

To derive an EM algorithm, the missing data must be considered. It seems obvious that the W_i are missing, and the treatment in section "A Deconvolution Problem" suggests that the Y_i are missing as well. So, the complete data is the random sample (J_i, W_i, Y_i) of the random variable (J, W, Y). This random variable lives in the statistical space $(\mathcal{M} \times \mathcal{Y} \times \mathcal{Y}, \mathcal{B}, \mathcal{Q})$, with $\mathcal{B}$ the σ-algebra generated by the sets $A \times B \times C$ with $A \subset \mathcal{M}$ and $B, C \in \mathcal{B}_{\mathcal{Y}}$. Finally, $\mathcal{Q}$ is the collection of probability measures on $\mathcal{B}$ that are absolutely continuous with respect to the product measure $\alpha \times \mu \times \mu$. The density of (J, W, Y) is then

$$f_{J,W,Y}(j,w,y) = k(w-y)\, f_Y(y)\, \mathbf{1}(w \in B_j), \quad j \in \mathcal{M},\ w,\ y \in \mathcal{Y}. \tag{76}$$

The complete maximum likelihood problem for estimating f_Y is now

$$\text{minimize} \quad -\frac{1}{n}\sum_{i=1}^{n} \log\{k(W_i - Y_i)\, f(Y_i)\} \quad \text{subject to} \quad f \in \mathcal{P}. \tag{77}$$

This is the same as the problem (58). However, here one conditions differently. At issue is the conditional expectation $\mathbb{E}\left[\log\{k(W-Y)\, f(Y)\} \,\middle|\, J\right]$. Observe that

$$f_{W,Y|J}(w,y|j) = \frac{f_{J,W,Y}(j,w,y)}{f_J(j)} = \frac{k(w-y)\, f_Y(y)\, \mathbf{1}(w \in B_j)}{[\mathcal{K} f_Y](B_j)}, \tag{78}$$

so that, replacing f_Y by some initial guess f_1, one finds that

$$\begin{aligned}&\mathbb{E}\left[\log\{k(W-Y)\, f(Y)\} \,\middle|\, J,\ f_1\right] \\ &\quad = \int_{\mathcal{Y}\times\mathcal{Y}} \log\{k(w-y)\, f(y)\}\, \frac{k(w-y)\, f_1(y)\, \mathbf{1}(w \in B_j)}{[\mathcal{K} f_1](B_j)}\, d\mu(w)\, d\mu(y).\end{aligned}$$

Now, using

$$\log\{k(w-y)\, f(y)\} = \log\{f(y)\} + (\text{a term not depending on } f),$$

one arrives at

$$\begin{aligned}&\mathbb{E}\left[\log\{k(W-Y)\, f(Y)\} \,\middle|\, J,\ f_1\right] \\ &\quad = \int_{\mathcal{Y}} \frac{f_1(y)\, k_j(y)}{[\mathcal{K} f_1](B_j)}\, \log\{f(y)\}\, d\mu(y) + \text{rem},\end{aligned} \tag{79}$$

where "rem" involves terms not depending on f, and

$$k_j(y) = \int_{B_j} k(w-y)\, d\mu(w), \quad j \in \mathcal{M}. \tag{80}$$

Note that then

$$[\mathcal{K} f_1](B_j) = \int_{\mathcal{Y}} k_j(y)\, f_1(y)\, d\mu(y). \tag{81}$$

So, the E-step of the EM algorithm leads to the problem

$$\text{minimize} \quad -\int_{\mathcal{Y}} f_2(y)\, \log f(y)\, d\mu(y) \quad \text{subject to} \quad f \in \mathcal{P}, \tag{82}$$

where (one would say: as always)

$$f_2(y) = f_1(y) \cdot \frac{1}{n} \sum_{i=1}^{n} \frac{k_{J_i}(y)}{[\mathcal{K} f_1](B_{J_i})}, \quad y \in \mathcal{Y}. \tag{83}$$

Since

$$\int_{\mathcal{Y}} f_1(y)\, d\mu(y) = \frac{1}{n} \sum_{i=1}^{n} \left\{ \int_{\mathcal{Y}} k_{J_i}(y)\, f(y)\, d\mu(y) \Big/ [\mathcal{K} f_1](B_j) \right\} = 1,$$

the solution of (82) is f_2. So, the iterative step of the EM algorithm is given by (83). Actually, the situation is a little sticky, since (83) involves the J_i. However, one may collect the J_i with equal values, so that then

$$N_j = \sum_{i=1}^{n} \mathbf{1}(j = J_i),$$

and so an equivalent definition of f_2 is

$$f_2(y) = f_1(y) \cdot \frac{1}{n} \sum_{j=1}^{m} \frac{N_j\, k_j(y)}{[\mathcal{K} f_1](B_j)} \quad y \in \mathcal{Y}. \tag{84}$$

The EM algorithm is then obtained by iterative applying of the EM-step (84).

Discretizing the EM algorithm: Note that a discretized EM algorithm may be derived by restricting the minimization in (74) to step functions on a partition $\{C_j\}_{j=1}^{\ell}$ of a set containing the support of Y. With $\mathcal{P}_m$ as in (64), the restricted maximum likelihood problem with binned data is

$$\text{minimize} \quad -\frac{1}{n} \sum_{j=1}^{\ell} N_j\, \log[\mathcal{K} f](C_j) \quad \text{subject to} \quad f \in \mathcal{P}_m. \tag{85}$$

One then derives the EM algorithm as in section "A Deconvolution Problem," leading to

$$x_j^{[k+1]} = x_j^{[k]} \cdot \frac{1}{n} \sum_{p=1}^{\ell} \frac{N_p\, k_{ip}}{\left(\sum_{q=1}^{\ell} a_{iq}\, x_q^{[k]} \right)}, \quad j = 1, 2, \ldots, \ell, \tag{86}$$

where for $p = 1, 2, \ldots, m$ and $q = 1, 2, \ldots, \ell$,

$$a_{pq} = \int_{C_p} \left\{ \int_{B_q} k(w-y)\, d\mu(y) \right\} d\mu(w).$$

The estimators for f_Y are then

$$f_k(y) = \sum_{j=1}^{\ell} x_j^{[k]} \mid C_j \mid^{-1} \mathbf{1}(y \in C_j), \quad y \in \mathcal{Y}. \tag{87}$$

Finite Mixtures of Unknown Distributions

The final simple case to be discussed is that of a mixture with a small number of densities belonging to some parametric family.

Consider a random variable Y in a statistical space $(\mathcal{Y}, \mathcal{B}_{\mathcal{Y}}, \mathcal{P})$, with

$$\mathcal{P} = \{ f(\cdot \mid x) : x \in \mathcal{X} \}, \tag{88}$$

a family of probability measures indexed by the (low-dimensional) parameter $x \in \mathcal{X}$. Assume that $\mathcal{P}$ is dominated by some measure P_∞ and that

$$\left[\frac{dP(\cdot \mid x)}{dP_\infty} \right](y) = f_{\mathcal{Y}}(y \mid x), \quad y \in \mathcal{Y}. \tag{89}$$

So, let Y be a random variable with density

$$f_Y(y) = \sum_{j=1}^{m} w_o(j) f_{\mathcal{Y}}\Big(y \Big| x_o(j)\Big), \quad y \in \mathcal{Y}. \tag{90}$$

Here, $w_o = (w_{o_1}, w_{o_2}, \ldots, w_{o_m}) \in V_m$, the space of probability vectors (see (43)), and $x_o = (x_{o,1}, x_{o,2}, \ldots, x_{o,m}) \in \mathcal{X}_m$, thus defining $\mathcal{X}_m$. (The notations $x_{o,j}$ and $x_o(j)$ are used interchangeably.)

Given a random sample $Y_1, Y_2, \ldots, Y_m$, the maximum likelihood problem for estimating w_o and x_o is then

$$\begin{aligned} \text{minimize} \quad & -\frac{1}{n} \sum_{i=1}^{n} \log \left(\sum_{j=1}^{m} w_j \, f_{\mathcal{Y}}(Y_i \mid x_j) \right) \\ \text{subject to} \quad & w \in V_m \,,\; x \in \mathcal{X}_m. \end{aligned} \tag{91}$$

To derive an EM algorithm, one must introduce the missing data. As in section "Mixtures of Known Densities," the random index $J \in \mathcal{M} = \{1, 2, \ldots, m\}$ would be a useful information because $f_{Y|J}(y \mid j) = f_{\mathcal{Y}}(y \mid x_{o,j})$.

So considering $(Y_1, J_1), (Y_2, J_2), \ldots, (Y_n, J_n)$ to be the complete data, the maximum likelihood problem is

$$\begin{aligned} &\text{minimize} \quad -\frac{1}{n}\sum_{i=1}^{n} \log\{w(J_i)\, f_{\mathcal{Y}}(Y_i|x(J_i))\} \\ &\text{subject to} \quad W \in V_m,\ x \in \mathcal{X}_m. \end{aligned} \tag{92}$$

Similar to the development in section "Mixtures of Known Densities," now with initial guesses $w^{[1]}$ and $x^{[1]}$, one obtains that

$$\begin{aligned} &\mathbb{E}\left[\log\{w(J)\, f_{\mathcal{Y}}(Y|x(j))\} \,\middle|\, Y, w^{[1]}, x^{[1]}\right] \\ &= \sum_{j\in\mathcal{M}} \frac{w^{[1]}(j)\, f_{\mathcal{Y}}\left(Y|x^{[1]}(j)\right)}{\left(\sum_{p=1}^{m} w^{[1]}(j)\, f_{\mathcal{Y}}\left(Y|x^{[1]}(j)\right)\right)} \log\{w(j)\, f_{\mathcal{Y}}\left(Y|x^{[1]}(j)\right)\}. \end{aligned}$$

It follows that

$$\begin{aligned} &\mathbb{E}\left[-\frac{1}{n}\sum_{i=1}^{n} \log\{w(J_i)\, f_{\mathcal{Y}}(Y_i|x(J_i),)\} \,\middle|\, Y_1, Y_2, \ldots, Y_n, w^{[1]}, x^{[1]}\right] \\ &= -\sum_{j\in\mathcal{M}} w^{[2]}(j)\, \log w(j) + \mathfrak{L}_n(x|x^{[1]}, w^{[1]}), \end{aligned} \tag{93}$$

where

$$w^{[2]}(j) = w^{[1]}(j) \cdot \frac{1}{n}\sum_{i=1}^{n} \frac{f_{\mathcal{Y}}(Y_i|x^{[1]}(j))}{\left(\sum_{p\in\mathcal{M}} w^{[1]}(p)\, f_{\mathcal{Y}}\left(Y_i|x^{[1]}(p)\right)\right)}, \quad j \in \mathcal{M}, \tag{94}$$

and

$$\mathfrak{L}_n\left(x|x^{[1]}, w^{[1]}\right) = -\frac{1}{n}\sum_{i=1}^{n} \frac{f_{\mathcal{Y}}(Y_i|x^{[1]}(j))\, \log f_{\mathcal{Y}}(Y_i|x(j))}{\left(\sum_{p\in\mathcal{M}} w^{[1]}(p)\, f_{\mathcal{Y}}\left(Y_i|x^{[1]}(p)\right)\right)}. \tag{95}$$

This is essentially the E-step of the EM algorithm. Note that the definition of $w^{[2]}$ is in the by-now-familiar form. For the M-step, one must solve

$$\begin{aligned} &\text{minimize} \quad -\sum_{j\in\mathcal{M}} w^{[2]}(j)\,\log w(j) + \mathfrak{L}_n\left(x|x^{[1]}, w^{[1]}\right) \\ &\text{subject to} \quad w \in V_m,\ x \in \mathcal{X}_m, \end{aligned} \tag{96}$$

and this nicely separates. The minimization over w gives $w = w^{[2]}$ as always, and $x = x^{[2]}$ is the solution of

$$\text{minimize} \quad \mathfrak{L}_n\left(x|x^{[1]}, w^{[1]}\right) \quad \text{subject to} \quad x \in \mathcal{X}_m. \tag{97}$$

Unfortunately, in general, there is no closed form solution of this problem.

There are numerous examples of this type of mixture problems. See, e.g., [58,79].

Empirical Bayes Estimation

There is another way of deriving EM algorithms for the mixture problem under consideration. Of course, that means one must introduce a different collection of missing data. The following development is from Eggermont and LaRiccia [39].

Starting from the beginning, consider the random variable Y with density $f_{\mathcal{Y}}(\,\cdot\,|x_o)$ with respect to P_∞. Given a random sample $Y_1, Y_2, \ldots, Y_n$ of the random variable Y, one wishes to estimate x_o. The maximum likelihood problem is

$$\text{minimize} \quad -\frac{1}{n}\sum_{i=1}^{n} \log f_{\mathcal{Y}}(Y_i|x) \quad \text{subject to} \quad x \in \mathcal{X}. \tag{98}$$

So, what is the missing data in this case? It was already alluded to: the missing information is x_o! In the so-called empirical Bayes approach, one considers x_o to be a random variable in the statistical space $(\mathcal{X}, \mathcal{B}_\mathcal{X}, \mathcal{T})$, with $\mathcal{T}$ a collection of probability measures on $\mathcal{B}_\mathcal{X}$, dominated by some measure T_∞. The complete data is the random sample $(Y_1, X_1), (Y_2, X_2), \ldots, (Y_n, X_n)$ of the random variable (Y, X), with density

$$f_{Y,X}(y, x) = f_X(x)\, f_{Y|X}(y|x) = f_X(x)\, f_{\mathcal{Y}}(y|x). \tag{99}$$

So, f_X is the marginal density of X, but instead of prescribing a prior distribution on X, one's task is to estimate this distribution without using prior information. The estimator of f_X will tell us whether one parameter $X = x_o$ suffices for all Y_i, viz., if the estimator of f_X has most of its mass near $x = x_o$ or if one has (mostly) a mixture with a few components, or indeed a continuous mixture.

Note that

$$f_Y(y) = \int_\mathcal{X} f_{\mathcal{Y}}(y|x)\, f_X(x)\, dT_\infty(x). \tag{100}$$

Defining the integral operator $\mathcal{K}$ by

$$[\mathcal{K}f](y) = \int_{\mathcal{X}} f_{\mathcal{Y}}(y|x)\, f(x)\, dT_{\infty}(x), \quad y \in \mathcal{Y}, \tag{101}$$

the maximum likelihood problem for estimating f_X is

$$\text{minimize} \quad -\frac{1}{n}\sum_{i=1}^{n} \log[\mathcal{K}f](Y_i) \quad \text{subject to} \quad f \in \mathcal{T}. \tag{102}$$

Note that the problem (102) is very much like the problem (57). Indeed, in very much the same way as in section "A Deconvolution Problem," one derives the EM algorithm

$$f_{k+1}(x) = f_k(x) \cdot \frac{1}{n}\sum_{i=1}^{n} \frac{f_{\mathcal{Y}}(Y_i|x)}{[\mathcal{K}f_k](Y_i)}, \quad x \in \mathcal{X}. \tag{103}$$

This pretty much exhausts the simple examples that lead to this "same old" EM algorithm.

5 Emission Tomography

Flavors of Emission Tomography

There are at least three flavors of emission tomography, viz., single photon emission tomography or SPECT (the C stands for "computerized"), positron emission tomography or PET, and time-of-flight PET (TOFPET). In all of these cases, given measurements of the emissions, the objective is to reconstruct the three-dimensional distribution of a radiopharmaceutical compound in the brain, giving insight into the metabolism in general and blood flow, in particular, in the brain.

In the single photon version, single photons are emitted in random locations in the brain and are detected (or not, as the case may be) by detectors situated around the head. In the positron version, single positrons are emitted in random locations. The positrons travel short distances until they are annihilated by single electrons, at which instances pairs of photons are created which fly off in nearly opposite random directions. The pairs of photons may then be detected by pairs of detectors. Thus, positron emission tomography amounts to double photon emission tomography. In the time-of-flight version of PET, the arrival times of the pairs of photons are recorded as well, which gives some information on the location of the emission. The time-of-flight version will not be considered further. Although the specifics are different, in their idealized form, the reconstruction problems for SPECT and PET are just about the same. For some of the details of the not-so-ideal circumstances in actual practice, see, e.g., [55, 97].

The Emission Tomography Experiment

The data collection experiment for emission tomography may be described as follows. Consider a three-dimensional Poisson random field living in an open ball $\Omega \in \mathbb{R}^d$ (with $d = 3$). Here, "events" (viz., the creation of a single or double photon) happen with spatial intensity per unit time denoted by $f_Z(z)$, $z \in \Omega$. This means that for any Borel subset C of Ω, the number $N(C, t, \delta t)$ of events that happen inside C during a time interval $(t, t + \delta t)$ does not depend on t and is a Poisson random variable with mean

$$\mathbb{E}[N(C, t, \delta t)] = \delta t \int_C f_Z(z)\, d\mu(z). \tag{104}$$

Moreover, if $C_1, C_2, \ldots, C_m$ are disjoint Borel subsets of Ω and $(t_i, t_i + \delta t_i)$, $i = 1, 2, \ldots, m$, denote arbitrary (deterministic) time intervals, then the counts $N(C_1, t_1, \delta t_1), N(C_2, t_2, \delta t_2), \ldots, N(C_m, t_m, \delta t_m)$, are independent. One may choose the unit of time ΔT in such a way that f_Z is the density of a probability measure with respect to Lebesgue measure on Ω. Then, in a time interval $(t, t + \lambda \Delta T)$, the number $N = N(\Omega, t, \lambda \Delta T)$ of events that occur throughout Ω is a Poisson random variable with mean

$$\lambda \int_\Omega f_Z(z)\, d\mu(z) = \lambda.$$

This may be written succinctly as

$$N \sim \text{Poisson}(\lambda). \tag{105}$$

For more on spatial Poisson processes, see, e.g., [24], section "The Shepp–Vardi EM Algorithm for PET."

Returning to the experiment, conditional on $N = n$, during the time interval $(0, \lambda \Delta T)$, one collects a random sample $Z_1, Z_2, \ldots, Z_n$ (of sample size n) of the random variable Z, the random location of an event, with density f_Z with respect to Lebesgue measure. The random variable Z lives in the statistical space $(\Omega, B_\Omega, \mathcal{P}_\Omega)$ with B_Ω the σ-algebra of Borel subsets of Ω and $\mathcal{P}_\Omega$ the collection of probability measures that are absolutely continuous with respect to Lebesgue measure. The events themselves are detected by detectors or pairs of detectors, denoted by $B_1, B_1, \ldots, B_m$, which one may view as disjoint subsets (or antipodal subsets) of a sphere surrounding Ω. For each event at a location Z_i, there is a random index $J \in \mathcal{M} = \{1, 2, \ldots, m\}$ such that the event is detected by the detector (pair) B_J. Thus, J lives in the statistical space $(\mathcal{M}, 2^{\mathcal{M}}, V_m)$; see (46). The random variable (Z, J) is absolutely continuous with respect to the product measure $\mu \times \alpha$, and its density is

$$f_{Z,J}(z, j) = f_{J|Z}(j\,|z)\, f_Z(z), \quad z \in \Omega\,,\ j \in \mathcal{M}, \tag{106}$$

with $f_{J|Z}$ determined by geometric considerations. See, e.g., [97]. Assume that this is known. Assuming that every event is detected, then $f_{J|Z}$ is a conditional density, so

$$\sum_{j=1}^{m} f_{J|Z}(j|z) = 1 \quad \text{for all } z. \tag{107}$$

Note that then

$$f_J(j) = \int_\Omega f_{J|Z}(j|z)\, f_Z(z)\, d\mu(z), \quad j \in \mathcal{M}. \tag{108}$$

So, conditional on $N = n$, one may pretend to have a random sample (Z_1, J_1), $(Z_2, J_2), \ldots, (Z_n, J_n)$ of the random variable (Z, J). This gives rise to the usual form of the actual data to wit the bin counts

$$N_j = \sum_{i=1}^{n} \mathbf{1}(J_i = j). \tag{109}$$

Continuing (and no longer conditioning on $N = n$), then $N_1, N_2, \ldots, N_m$ are independent Poisson random variables with $\mathbb{E}[N_j] = [\mathcal{K} f_Z](j)$,

$$N_j \sim \text{Poisson}\left(\lambda\, [\mathcal{K} f_Z](j)\right), \quad j \in \mathcal{M}, \tag{110}$$

where $\mathcal{K} : L^1(\Omega, d\mu) \longrightarrow L^1(\mathcal{M}, d\alpha)$ is defined by

$$[\mathcal{K}\varphi](j) = \int_\Omega f_{J|Z}(j|z)\, \varphi(z)\, d\mu(z), \quad j \in \mathcal{M}. \tag{111}$$

This concludes the description of the ideal emission tomography experiment. In reality, quite a few extra things need to be taken into account, such as the attenuation of photons by tissue and bone in the head; see, e.g., [55]. For the treatment of background noise, see, e.g., [42] and references therein.

The Shepp–Vardi EM Algorithm for PET

After these preparations, the maximum likelihood problem for estimating f_Z may be formulated. The observed data are the count data $N_1, N_2, \ldots, N_m$, which leads to the problem

$$\text{minimize} \quad -\frac{1}{N} \sum_{j=1}^{m} N_j \log[\mathcal{K} f](j) \quad \text{subject to} \quad f \in \mathcal{P}_\Omega. \tag{112}$$

Alternatively, and this is actually more convenient, one may view the total number of detected events N and $J_1, J_1, \ldots, J_N$ as the actual data, which gives

$$\text{minimize} \quad -\frac{1}{N}\sum_{i=1}^{N} \log[\mathcal{K}f](J_i) \quad \text{subject to} \quad f \in \mathcal{P}_\Omega. \tag{113}$$

In view of Remark 2 following Remark (75), these two problems are equivalent.

So far, nothing has been said about approximating/representing the estimators in terms of pixels or voxels. Let $\{C_p\}_{p=1}^{\ell}$ be a partition of Ω, and let $\mathcal{P}_\ell \subset \mathcal{P}_\Omega$ be the space of step functions that are constant on each C_p. Thus, with V_ℓ defined as in (43),

$$\mathcal{P}_\ell = \left\{ \sum_{p=1}^{\ell} x_p\, b_p(\,\cdot\,) \,\middle|\, x \in V_\ell \right\}, \tag{114}$$

where $$b_p(z) = |\,C_p\,|^{-1}\,\mathbf{1}(z \in C_p), \quad z \in \Omega.$$

(Note however that one may take other basis functions.) The discretized maximum likelihood problem is then obtained by restriction

$$\text{minimize} \quad -\frac{1}{N}\sum_{i=1}^{N} \log[\mathcal{K}f](J_i) \quad \text{subject to} \quad f \in \mathcal{P}_\ell. \tag{115}$$

From the description of the experiment, it is clear that the missing data are the Z_i, so the complete data set is $(Z_1, J_1), (Z_2, J_2), \ldots, (Z_N, J_N)$, a random sample (of random sample size N) of the random variable (Z, J). The joint density of $N, (Z_1, J_1), (Z_2, J_2), \ldots, (Z_N, J_N)$ is then

$$\frac{\lambda^n}{n\,!}\,\mathrm{e}^{-\lambda} \prod_{i=1}^{n} \Big\{ f_{J|Z}(j_i|z_i)\, f_Z(z_i) \Big\}, \tag{116}$$

so that the complete maximum likelihood problem is

$$\text{minimize} \quad -\frac{1}{N}\sum_{i=1}^{N} \log\Big\{ f_{J|Z}(J_i|Z_i)\, f(Z_i) \Big\} \quad \text{subject to} \quad f \in \mathcal{P}_\ell. \tag{117}$$

In the objective function, the terms corresponding to the Poisson distribution of N have been omitted, and the scaling $1/N$ was applied.

For the E-step of the EM algorithm, consider the computation of

$$\mathbb{E}\Big[-\log\Big\{ f_{J|Z}(j|z)\, f(Z) \Big\} \,\Big|\, J,\ f_1 \Big], \tag{118}$$

assuming the approximation f_1 to f_Z. Since

$$f_{Z|J}(z|j) = \frac{f_{J|Z}(j|z)\, f_Z(z)}{f_J(j)}, \tag{119}$$

this gives for the conditional expectation (118) the expression

$$-\int_\Omega \frac{f_{J|Z}(j|z)\, f_Z(z)}{f_J(j)} \log f(z)\, d\mu(z), \tag{120}$$

where the contribution involving the known $f_{J|Z}$ may be ignored, since it does not depend on f.

Consequently, the E-step leads to the problem

$$\text{minimize} \quad -\int_\Omega f_2(z) \log f(z)\, d\mu(z) \quad \text{subject to} \quad f \in \mathcal{P}_\ell, \tag{121}$$

where

$$f_2(z) = f_1(z) \cdot \frac{1}{N} \sum_{i=1}^{N} \frac{f_{J|Z}(J_i|z)}{\left(\int_\Omega f_{J|Z}(J_i|s)\, f_1(s)\, d\mu(s)\right)}. \tag{122}$$

Note that f_2 is a density.

In terms of the representation of elements in $\mathcal{P}_\ell$,

$$f(z) = \sum_{p=1}^{\ell} x_p\, a_p(z), \tag{123}$$

with $a_p(z) = |\,C_p\,|^{-1}\, \mathbf{1}(z \in C_p)$ as in (62), and likewise for f_1 and f_2, this leads to

$$\begin{aligned} \int_\Omega f_2(z) \log f(z)\, d\mu(z) &= \sum_{p=1}^{\ell} x_p^{[2]} \log \left\{x_p\, |\,C_p\,|^{-1}\right\} \\ &= \sum_{p=1}^{\ell} x_p^{[2]} \log x_p - \log |\,C_p\,| \sum_{p=1}^{\ell} x_p^{[2]} \\ &= \sum_{p=1}^{\ell} x_p^{[2]} \log x_p - \log |\,C_p\,|, \end{aligned} \tag{124}$$

where

$$x_p^{[2]} = x_p^{[1]} \cdot \frac{1}{N} \sum_{i=1}^{N} \frac{a(J_i, p)}{\left(\sum_{q=1}^{\ell} a(J_i, p)\, x_p^{[1]} \right)}, \tag{125}$$

with

$$a(j, p) = \int_{\mathbb{R}^d} f_{J|Z}(j|z)\, a_p(z)\, d\mu(z) = \int_{C_p} f_{J|Z}(j|z)\, d\mu(z). \tag{126}$$

Note that $\sum_{p=1}^{\ell} x_p^{[2]} = 1$ and, by (107), that

$$\sum_{j=1}^{m} a(j, p) = 1 \quad \text{for all } p. \tag{127}$$

So, the E-step gives

$$\text{minimize} \quad -\sum_{p=1}^{\ell} x_p^{[2]} \log x_p \quad \text{subject to} \quad x \in V_\ell. \tag{128}$$

In section "Mixtures of Known Densities," it was already shown that the solution of the problem (128) is given by $x = x^{[2]}$. So (125) is the iterative step of the EM algorithm. Of course, using Remark 2, the iterative step for $x^{[2]}$ may be rewritten in terms of the bin counts as

$$x_p^{[2]} = x_p^{[1]} \cdot \frac{1}{N} \sum_{j=1}^{m} \frac{N_j\, a(j, p)}{\left(\sum_{q=1}^{\ell} a(j, p)\, x_p^{[1]} \right)}. \tag{129}$$

Observe again the similarity with the EM algorithms for the simple examples in Sect. 4.

Remark 3. The problem (115) is not really discretized. The actual discretized problem is

$$\text{minimize} \quad -\frac{1}{N}\sum_{j=1}^{m} N_j \log[Ax]_j + \sum_{p=1}^{\ell} x_p \quad \text{subject to} \quad x \in V_\ell, \tag{130}$$

with V_ℓ given by (43), and $A \in \mathbb{R}^{m\times\ell}$ has components $a(j,p)$ given by (126). This uses Remark 2.

Remark 4. To finish, note that the original derivation by Shepp and Vardi [88] involved the missing data $M(j,p)$, the number of events in each "cell" of Ω that contribute to the counts N_j,

$$M(j,p) = \sum_{i=1}^{m} \mathbf{1}(J_i = j)\,\mathbf{1}(Z_i \in C_p), \quad p = 1,2,\ldots,\ell.$$

This calls for a rather complicated relation between the $M(j,p)$ and the N_j. In particular, one does not have random samples of the appropriate random variables. It gets much simplified if one introduces the random variables $I_1, I_2, \ldots, I_m$ which indicate to what "cell" the event Z_i belongs to. So, $I_i = p$ if

$$\mathbf{1}(Z_i \in C_p) = 1.$$

Then, for the complete data, one gets back to considering random samples of random variables, viz., (J, I). This would provide for an alternative approach to discretization but would lead to the same EM algorithm. This is essentially the "list mode" approach of Parra and Barrett [77]. See also [10].

Prehistory of the Shepp–Vardi EM Algorithm

The earliest reference to maximum likelihood estimation in emission tomography is the aforementioned paper by Rockmore and Macovski [82]. In astronomy, an early reference is Lucy [70]. The EM algorithm for these maximum likelihood problems was introduced by Shepp and Vardi [88] and independently by Lange and Carson [64]. See also [20] for a completely different setting. The EM algorithm for SPECT is essentially the same; see, e.g., [60] and references therein.

The algorithm itself may be viewed as a method for approximately solving the integral equation with moment discretization

$$[\mathcal{K}f](j) = \frac{N_j}{N}, \quad j = 1,2,\ldots,m, \tag{131}$$

with $\mathcal{K}$ as in (111). In particular, this may be applied to Fredholm integral equations of the first kind. As such it was independently discovered in various settings many times over, by Tarasko [93] and Kondor [61] in Physics, by Richardson [81] and Lucy [70] in Astronomy, and perhaps other authors. It is interesting to note that both Richardson [81] and Lucy [70] derive the algorithm based on probabilistic

considerations involving Bayes' theorem, as in (119). For more on the integral equations aspect, see also [74].

6 Electron Microscopy

In this section, a recent application of EM algorithms at the bleeding edge of science is considered. As far as the EM algorithm is concerned, the foundation is far from complete, whether it be practical or theoretical.

Imaging Macromolecular Assemblies

Structural biologists are interested in the shape of biological objects at the macromolecular level. The tail of the T4 bacteriophage is a famous example. Such objects are referred to as macromolecular assemblies. To view objects that are this tiny, electron microscopy seems to be the only tool available. Its use in structural biology goes back to DeRosier and Klug; see [21]. Ideally, one would like to take a single tail of the T4 bacteriophage, say, obtain electron micrographs (projections) from many directions, and reconstruct the three-dimensional structure of the tail. Unfortunately, the bombardment with electrons destroys the object, so that only one projection can be taken. The biologists have found a way around this, but it comes at a price. Very roughly speaking, many tails are isolated and suspended in a thin layer of water, which is then rapidly cooled to below freezing. This results in vitreous water, with the tails suspended in it *but randomly located and oriented*. A single electron micrograph of this layer is then taken. This is equivalent to taking projections of a single tail in many different directions, corresponding to the random orientations. For a precise description and analysis of the procedure, see [44]. Now, the price one pays is that one has many projections of the tail but in random unknown directions. Since these random directions may be viewed as missing data, it is clear that EM algorithms may be used. This was first realized by Scheres et al. [85]. A complication is that the objects can appear in conformational states, which means that one has a mixture of finitely many (different) objects. Another complication is that the signal-to-noise ratio is typically quite small, in the 10 % range.

The Maximum Likelihood Problem

Mathematically, following Scheres et al. [83, 84], the setup may be described as follows. Each object in the thin layer may be considered as being randomly chosen from a finite collection $x_1^o, x_2^o, \ldots, x_\kappa^o$ of κ objects. Its position and orientation in the thin layer is described by five real-valued parameters: two location parameters and three Eulerian angles describing its orientation. Denote them by Θ and the set of all possible Θ by Ξ. The problem of finding the location parameters is referred to as the problem of alignment. For low signal-to-noise ratios, maximum likelihood methods seem to be preferable [89].

So, for a random object, one observes the projection in the form of a discretized image Y,

$$Y = C * R_\Theta\, x_K^o + \varepsilon. \tag{132}$$

Here, K is the random index into the collection of possible objects, $R_\Theta x_K$ is the projection data of the object in the "direction" Θ, C is the (known) contrast transfer function (due to the experimental setup), and $*$ denotes the two-dimensional discretized convolution operation. Finally, ε is the noise, assumed to be normal and isotropic, i.e., the components of ε are jointly normal and the components of the variance–covariance matrix V_o satisfy

$$\mathbb{E}[\varepsilon_{p,q}\,\varepsilon_{r,s}] = \Sigma_{p-r,q-s}, \tag{133}$$

where $\Sigma_{p-r,q-s}$ is a function of $(p-r)^2+(q-s)^2$, the (squared) Euclidean distance, only. In terms of the two-dimensional discrete Fourier transform, this means that (ignoring boundary effects)

$$\mathbb{E}\left[\hat{\varepsilon}_{P,Q}\,\overline{\hat{\varepsilon}_{R,S}}\right] = \left[\sigma_o^2\right]_{P,Q} \quad \text{for} \quad P = R \quad \text{and} \quad Q = S, \tag{134}$$

and $= 0$ otherwise. Moreover, $\sigma^2_{P,Q}$ is rotationally symmetric, i.e., it is a function of $P^2 + Q^2$. Unfortunately, $\sigma^2_{P,Q}$ is unknown; it must be estimated. Moreover, σ^2 varies with Y.

So, the distribution of Y conditional on $K = k$ and $\Theta = \theta$ is given by

$$f_{Y|\Theta,K}\left(y|\theta, x_k^o\right) = \frac{\exp\left(-\frac{1}{2}\,\|\,C * R_\theta\, x_k^o - y\,\|^2_{V_o}\right)}{(2\pi)^{N/2}\,\det(V_o)}, \tag{135}$$

where N is the size of Y (or y) and $\|\,z\,\|^2_V = z^T V^{-1} z$. In terms of Fourier transforms, this reads as

$$\begin{aligned} \|\,C * R_\theta\, x_k^o - y\,\|^2_{V_o} &= \sum_{P,Q} \left[\sigma_o^{-2}\right]_{P,Q} \left|\left[\hat{C}\right]_{P,Q} \left[R_\theta\, x_k^o\right]^\wedge_{P,Q} - [\hat{y}]_{P,Q}\right|^2, \\ \log\det(V_o) &= \sum_{P,Q} \log\left[\sigma_o^2\right]_{P,Q}. \end{aligned} \tag{136}$$

Now introduce the state of the system Y one wishes to estimate and the initial state of one's understanding of the system,

$$S = \left\{\varpi^o, f_{\Theta|K}, x^o, \sigma_o\right\} \quad \text{and} \quad S^{[1]} = \left\{\, bar\varpi^{[1]}, \varphi^{[1]}, x^{[1]}, \sigma^{[1]}\right\}, \tag{137}$$

where $\varpi_k^o = \mathbb{P}[K = k]$, $f_{\Theta|K}$ is the density of Θ conditional on K, and x^o and σ_o are as before. The current understanding of the system is comprised of one's best guesses so far for the true system.

The distribution of Y may be expressed as

$$f_Y(y) = \sum_{k=1}^{\kappa} \varpi_k^o \int_{\Xi} f_{Y|\Theta,K}(y|\theta, x_k)\, f_{\Theta|K}(\theta|k)\, d\nu(\theta), \tag{138}$$

where $\nu = \mu \times \omega$, with μ the Lebesgue measure on $\mathbb{R}^2$ and ω the surface measure on the sphere in $\mathbb{R}^3$. Since it is reasonable to assume that the random location parameters are independent of the random orientation, then

$$f_\Theta(\theta_1, \theta_2, \theta_3, \theta_4, \theta_5) = g(\theta_1, \theta_2)\, h(\theta_3, \theta_4, \theta_5), \tag{139}$$

for appropriate densities g and h. This reduces the actual dimension of the problem, but for notational ease, such a specialization will not be made.

Given a random sample $Y_1, Y_2, \ldots, Y_n$, of Y, the maximum likelihood problem for estimating the unknown objects $x_1, x_2, \ldots, x_\kappa$ may then be formulated:

$$\text{minimize } -\frac{1}{n} \sum_{i=1}^{n} \log \left(\sum_{k=1}^{\kappa} \varpi_k \int_{\Xi} \varphi(\Theta_i|k)\, f_{Y|K,\Theta}(Y_i|x_k, \theta)\, d\nu(\theta) \right). \tag{140}$$

The unknowns are the probability vector ϖ, the densities $\varphi(\theta|k)$ (keep (139) in mind), the unknown objects $x_1, x_2, \ldots, x_\kappa$, and the variances $\left[\sigma_i^2\right]_{P,Q}$. Note that there is some similarity with the empirical Bayes problem of section "Empirical Bayes Estimation."

The EM Algorithm, up to a Point

Obviously, the goal is to derive an EM algorithm for the solution of (139), but the final algorithm is not quite the real thing. It is clear that the missing data for each observed projection Y_i consists of the orientation, denoted by Θ_i, and which kind of object one is looking at, encoded in the index K_i. The σ_i^2 and the objects x_k are considered as parameters. So, the complete data set is (Y_i, Θ_i, K_i), $i = 1, 2, \ldots, n$, and the complete maximum likelihood problem is then to minimize

$$\Lambda_n(S) \stackrel{\text{def}}{=} -\frac{1}{n} \sum_{i=1}^{n} \log \left(\varpi_{K_i}\, \varphi(\Theta_i|K_i)\, f_{Y|\Theta,K}(Y_i|\Theta_i, x_{K_i}, \sigma_i) \right) \tag{141}$$

over all probability vectors ϖ, all densities $\varphi(\cdot|k)$, all variance matrices σ_i^2, and all $x_1, x_2, \ldots, x_\kappa$. However, recall that the $\varphi(\cdot|k)$ have a simple structure.

For the E-step, the conditional expectation $\mathbb{E}[\Lambda_n(S)\,|\,Y]$ is needed. By Bayes' rule, the distribution of (K, Θ) conditional on Y is described by

$$\mathbb{P}[K=k|Y=y]\, f_{\Theta|Y,K}(\theta|y,x_k) = \frac{\varpi_k^o\, f_{\Theta|K}(\theta|k)\, f_{Y|\Theta,K}(y|\theta,x_k,\sigma)}{f_Y(y)}.$$

So, with the current state $S^{[1]}$, and setting $\mathbb{Y}_n = \{Y_1, Y_2, \ldots, Y_n\}$, one gets

$$\mathbb{E}[\Lambda_n(S)\,|\mathbb{Y}_n, S^{[1]}] = \sum_{k=1}^{\kappa} \int_\Xi g_k^{[2]}\left(\theta, Y_i\right) \log\left(\varpi_k\, \varphi(\theta|k)\, f_{Y|\Theta,K}\left(Y_i|\theta, x_k^{[1]}, \sigma_i^{[1]}\right)\right) d\nu(\theta), \tag{142}$$

where

$$g_k^{[2]}(\theta, Y_i) = \frac{\varpi_k^{[1]}\, \varphi^{[1]}(\theta|k)\, f_{Y|\Theta,K}\left(Y_i|\theta, x_k^{[1]}, \sigma_i^{[1]}\right)}{f_Y^{[1]}(Y_i)}, \quad \text{with} \tag{143}$$

$$f_Y^{[1]}(y) = \sum_{k=1}^{\kappa} \varpi_k^{[1]} \int_\Xi \varphi^{[1]}(\theta|k)\, f_{Y|\Theta,K}\left(y|\theta^{[1]}, x_k^{[1]}\right) d\nu(\theta). \tag{144}$$

This completes the E-step.

The M-step deals with the minimization of $\mathbb{E}[\Lambda_n(S)\,|\mathbb{Y}_n, S^{[1]}]$ over S. This separates into three problems. First, estimating ϖ may be done by solving

$$\begin{aligned} &\text{minimize} \quad -\sum_{k=1}^{\kappa} (\log \varpi_k) \int_\Xi h_k^{[2]}(\theta)\, d\nu(\theta) \\ &\text{subject to} \quad \varpi \text{ is a probability vector,} \end{aligned} \tag{145}$$

where

$$h_k^{[2]}(\theta) = \frac{1}{n} \sum_{i=1}^{n} g_k^{[2]}(\theta|Y_i). \tag{146}$$

One verifies that the solution is $\varpi = \varpi^{[2]}$,

$$\varpi_k^{[2]} = \int_\Xi h_k^{[2]}(\theta)\, d\nu(\theta), \quad k = 1, 2, \ldots, \kappa. \tag{147}$$

Second, estimating the $\varphi(\,\cdot\,|k)$ may be done by solving

$$\begin{aligned} &\text{minimize} \quad -\sum_{k=1}^{\kappa} \int_\Xi h_k^{[2]}(\theta, Y_i)\, \log \varphi(\theta|k)\, d\nu(\theta) \\ &\text{subject to} \quad \varphi(\,\cdot\,|k) \text{ is a pdf, } k = 1, 2, \ldots, \kappa. \end{aligned} \tag{148}$$

This separates into κ minimization problems for the $\varphi(\,\cdot\,|k)$. One verifies that the solutions are for $k = 1, 2, \ldots, \kappa$,

$$\varphi^{[2]}(\theta|k) = \varphi^{[1]}(\theta|k) \cdot \frac{1}{n} \sum_{i=1}^{n} \frac{f_{Y|\Theta,K}\left(Y_i|\theta, x_k^{[1]}, \sigma_i^{[1]}\right)}{f_{Y|K}^{[1]}(Y_i|k)}, \tag{149}$$

where

$$f_{Y|K}^{[1]}(Y_i|k) = \int_{\Xi} \varphi^{[1]}(\theta|k)\, f_{Y|\Theta,K}\left(Y_i|\theta, x_k^{[1]}, \sigma_i^{[1]}\right) d\nu(\theta). \tag{150}$$

Note that (147) and (148) are again multiplicative algorithms.

Third and last, one must estimate $f_{Y|\Theta,K}$, which boils down to estimating the x_k and σ_i. The problem is to minimize

$$-\sum_{k=1}^{\kappa} \int_{\Xi} \frac{1}{n} \sum_{i=1}^{n} g_k^{[2]}(\theta, Y_i)\, \log f_{Y|\Theta,K}(Y_i|\theta, x_k, \sigma_i)\, d\nu(\theta) \tag{151}$$

over x_k and σ_i. Since $\log f_{Y|\Theta,K}(Y_i|\theta, x_k, \sigma_i)$ equals

$$\sum_{P,Q} \left\{ \frac{\left| \hat{C}_{P,Q} \cdot [(R_\theta\, x_k)^\wedge]_{P,Q} - [\hat{Y}_i]_{P,Q} \right|^2}{\left[2\sigma_i^2\right]_{P,Q}} + \log\left[\sigma_i^2\right]_{P,Q} \right\},$$

here too the minimization problems separate. This may be solved for each x_k by minimizing

$$\mathrm{WLS}_k(x_k) = \int_{\Xi} \frac{1}{n} \sum_{i=1}^{n} g_k^{[2]}(\theta, Y_i) \left\{ \frac{\left| \hat{C}_{P,Q} \cdot [(R_\theta\, x_k)^\wedge]_{P,Q} - \left[\hat{Y}_i\right]_{P,Q} \right|^2}{\left[2\sigma_i^2\right]_{P,Q}} \right\} d\nu(\theta).$$

Denoting the minimizing x_k by $x_k^{[2]}$, then the new σ_i is $\sigma_i = \sigma_i^{[2]}$ with

$$\begin{aligned} \left[\sigma_i^{[2]}\right]_{P,Q} = &\sum_{k=1}^{\kappa} \int_{\Xi} \frac{1}{n} \sum_{i=1}^{n} g_k^{[2]}(\theta, Y_i) \\ &\times \left\{ \left| \hat{C}_{P,Q} \cdot \left[\left(R_\theta\, x_k^{[2]}\right)^\wedge \right]_{P,Q} - \left[\hat{Y}_i\right]_{P,Q} \right|^2 \right\} d\nu(\theta). \end{aligned}$$

The Ill-Posed Weighted Least-Squares Problem

Up to this point, the M-step has been carried out exactly. The last part is to minimize $\mathrm{WLS}_k(x_k)$. This is a weighted least-squares problem, which is nice, but it is ill-posed (or ill conditioned after discretization), which implies that one cannot and should not solve it exactly. In fact, Scheres et al. [84] employ a Wiener filter to stably implement the "exact" deconvolution procedure

$$\left[\left(R_\theta\, x_k^{[2]}\right)^\wedge\right]_{P,Q} = \frac{[\hat{Y}_i]_{P,Q}}{\hat{C}_{P,Q}} \quad \text{for all } P, Q,$$

as in Penczek et al. [78], compared with Byrne and Fiddy [14], after which a weighted least-squares version of ART (WLSART) is applied.

It should be observed that these problems are very large and are computationally very expensive. It is clear that there is much room for algorithmic development, but the present discussion ends here.

7 Regularization in Emission Tomography

The Need for Regularization

It is clear that the simple deconvolution problem (57) and the more complicated PET problem (112) are ill-posed, let alone the electron microscopy problem of Sect. 6. As already observed, in the PET problem (112), one is trying to solve the compact operator equation with moment discretization

$$[\mathcal{K}f](j) = b_j, \quad j = 1, 2, \ldots, m, \tag{152}$$

where $b_j = N_j/N$. The possible nonexistence of solutions is dealt with by considering the maximum likelihood problem, which may be reformulated as

$$\text{minimize} \quad \mathrm{KL}(b, \mathfrak{K}f) \quad \text{subject to} \quad f \in \mathcal{P}, \tag{153}$$

where KL is the discrete Kullback–Leibler divergence (see (54)) and

$$\mathfrak{K}f = ([\mathcal{K}f](1), [\mathcal{K}f](2), \ldots, [\mathcal{K}f](m))^T.$$

The problem (153) is similar to a least-squares problem. However, as in the case of the least-squares approach to compact operator equations, this still does not take care of the ill-posedness of the problem. So, the problem (153) must be regularized.

The standard and practically the most often used method is to use the EM algorithm and stop the algorithm at some appropriate point in the iteration. See, e.g., [69] for practical aspects and [80] and [51] for some theoretical results. The alternative is

essentially Tikhonov regularization of the negative log-likelihood, which also comes in the guises of Bayesian or maximum a posteriori (MAP) likelihood estimation, Gibbs smoothing, or just roughness penalization. See [36, 46, 54, 63, 73]. A new twist is the use of total variation regularization and nonlinear diffusion filtering in connection with maximum likelihood estimation and EM algorithms (see, e.g., [3, 6, 31, 87, 100]), but unfortunately, this will not be discussed further.

In many cases, the E-step of the EM algorithm may be carried out explicitly, but not so for the M-step. Here, some obvious modifications of the EM algorithm or extraneous iterative methods must be introduced. However, a few examples of explicit honest-to-goodness EM algorithms for regularized maximum likelihood problems are discussed: the NEMS modification of the EMS method of Silverman et al. [90] and two EM algorithms for Good's roughness penalization.

Smoothed EM Algorithms

In this section, the discussion centers on the EMS algorithm of Silverman et al. [90] and the nonlinearly smoothed NEMS variant of Eggermont and LaRiccia [36] in the context of the deconvolution problem of section "A Deconvolution Problem." Silverman et al. [90] realized the necessity for regularization of the maximum likelihood problem in that the EM algorithm produces increasingly rougher estimators. Initially, this is good since one typically starts out with a uniform estimator and more features of the signal appear. However, as the iteration progresses, the estimator becomes increasingly nonsmooth, giving rise to spurious features. But Silverman et al. [90] figured they knew how to fix the nonsmoothness: Add a smoothing step to the EM algorithm.

So, let $\mathcal{S}_h$ be a smoothing operator in the form

$$[\mathcal{S}_h f](y) = \int_{\mathbb{R}^d} S_h(y-z)\, f(z)\, d\mu(z), \quad y \in \mathbb{R}^d, \tag{154}$$

where $S_h(z) = h^{-d}\, S(h^{-1}z)$ for some bounded, continuous, symmetric pdf $S \in L^1(\mathbb{R}^d)$, possibly with compact support. The EMS algorithm then takes the form

$$\begin{aligned} f_{k+1/2}(z) &= f_k(z) \cdot \frac{1}{n} \sum_{i=1}^{n} \frac{k(Y_i - z)}{[\mathcal{K} f_k](Y_i)}, \\ f_{k+1} &= \mathcal{S}_h f_{k+1/2}. \end{aligned} \tag{155}$$

So the general step of the EMS algorithm is one step of the EM algorithm followed by one smoothing step.

Silverman et al. [90] apply this algorithm to the simple problem of stereology (a integral equation on a compact interval) and to positron emission tomography. In both cases, it seems to work quite well. Of course, the question is whether the algorithm (155) converges and, if so, what it converges to. Regarding the first

question, see [65, 99]. Characterizing the limit is not so easy, e.g., if one has a fixed point of the iteration, does one then have a point where the gradient of some log-likelihood-like function vanishes?

In retrospect, it is clear that adding a smoothing step to the EM algorithm is a fundamentally sound idea, but the way it is implemented is not "right." Indeed, in view of the multiplicative character of the EM algorithm, it seems that multiplicative smoothing is called for. So, with $\mathcal{S}_h$ as before but with $S_h(z) \geqslant 0$ everywhere, define the nonlinear smoothing operator $\mathcal{N}$ on nonnegative functions f by

$$[\mathcal{N}(f)](y) = \exp\left([\mathcal{S}_h(\log f)](y)\right), \quad y \in \mathbb{R}^d. \tag{156}$$

Note that by convexity, $[\mathcal{N}(f)](y) \leqslant [\mathcal{S}_h f](y)$, so that $\mathcal{N}(f)$ is always well defined. One verifies that $\mathcal{N}$ performs multiplicative smoothing, i.e.,

$$\mathcal{N}(f \cdot g) = \mathcal{N}(f) \cdot \mathcal{N}(g), \tag{157}$$

where the dot means pointwise multiplication: $[f \cdot g](y) = f(y)g(y)$ for all y. It now turns out that the smoothed maximum likelihood problem

$$\text{minimize} \quad -\sum_{i=1}^{n} \log[\mathcal{K}\mathcal{N}(f)](Y_i) \quad \text{subject to} \quad f \in \mathcal{P} \tag{158}$$

admits the EM algorithm

$$\begin{aligned} f_{k+1/3} &= \mathcal{N}(f_k), \\ f_{k+2/3}(z) &= f_{k+1/3}(z) \cdot \frac{1}{n} \sum_{i=1}^{n} \frac{k(Y_i - z)}{[\mathcal{K} f_{k+1/3}](Y_i)}, \\ f_{k+1} &= S_h f_{k+2/3}, \end{aligned} \tag{159}$$

see [36]. In addition, the problem (158) has a solution and it is unique, and the algorithm (159) converges to this solution in the Kullback–Leibler sense. See section "Monotonicity of the Smoothed EM Algorithm."

The algorithm (159) is referred to as the NEMS algorithm: The general step consists of a nonlinear smoothing step, one step of the original EM algorithm, and a final (linear) smoothing step. The practical performance on the toy stereology problem is just about indistinguishable from the NEMS algorithm except that with the same smoothing operator $\mathcal{S}_h$, the NEMS algorithm does about twice the smoothing of the EMS algorithm. Note that the question about the proper choice of the smoothing operator (or smoothing matrix in the discrete case) arises. This is in effect a question about the selection of the regularization parameter in ill-posed problems. Unfortunately, this is not addressed in this chapter.

Good's Roughness Penalization

Good's roughness penalization of the deconvolution problem is a particular form of Tikhonov regularization. The roughness penalty function of Good [47] is

$$\Phi(f) = \frac{1}{4} \int_{\mathbb{R}^d} \frac{|\nabla f(z)|^2}{f(z)}\, d\mu(z). \tag{160}$$

(The factor $\frac{1}{4}$ is for convenience only.) The maximum penalized likelihood problem is then

$$\begin{aligned} &\text{minimize} \quad -\frac{1}{n}\sum_{i=1}^{n} \log[\mathcal{K}f](Y_i) + \int_{\mathbb{R}^d} f(z)\, d\mu(z) + h^2\, \Phi(f) \\ &\text{subject to} \quad f \in \mathcal{P}. \end{aligned} \tag{161}$$

One can now perform the E-step as in section "A Deconvolution Problem" to arrive at the problem

$$\begin{aligned} &\text{minimize} \quad -\int_{\mathbb{R}^d} f_2(y)\, \log f(y)\, d\mu(y) + \int_{\mathbb{R}^d} f(y)\, dy + h^2\, \Phi(f) \\ &\text{subject to} \quad f \in \mathcal{P}, \end{aligned} \tag{162}$$

where

$$f_2(y) = f_1(y) \cdot \frac{1}{n}\sum_{i=1}^{n} \frac{k(W_i - y)}{[\mathcal{K} f_1](W_i)}, \quad y \in \mathbb{R}^d. \tag{163}$$

At this stage, the change of variable $u = \sqrt{f}$ is obviously(?) useful. The problem then becomes

$$\begin{aligned} &\text{minimize} \quad -2\int_{\mathbb{R}^d} f_2(y)\, \log u(y)\, d\mu(y) + \|u\|^2 + h^2\, \|\nabla u\|^2 \\ &\text{subject to} \quad u \in L^2(\mathbb{R}^d),\ u \geqslant 0, \end{aligned} \tag{164}$$

where $\|\cdot\|$ denotes the $L^2(\mathbb{R})$ norm. Here, it is convenient to drop the constraint $\|u\| = 1$. Note that (164) is a convex minimization problem. The Euler equations are given by the boundary value problem

$$\begin{aligned} &-h^2\, \Delta u + u = \frac{f_2}{u} \quad \text{in} \quad \mathbb{R}^d, \\ &\nabla u(y) \longrightarrow 0 \quad \text{for} \quad |y| \longrightarrow \infty, \end{aligned} \tag{165}$$

where u is nonnegative. The M-step amounts to solving the boundary value problem.

The resulting algorithm converges, by arguments similar to those for the related discrete case of the next section. See section "Monotonicity for Exact Gibbs Smoothing."

For the positron emission tomography problem, Miller and Roysam [73] arrived at the analogue of this equation and solved the boundary value problem by finite differences, using Jacobi's method on a massively parallel computer. Of course, other methods come to mind.

Another EM algorithm: There is another way to proceed. With the change of variable $u = \sqrt{f}$ as before, the objective function in (161) becomes

$$-\frac{1}{n}\sum_{i=1}^{n}\log[\mathcal{K}(u^2)](Y_i) + \|\,u\,\|^2 + h^2\,\|\,\nabla u\,\|^2. \tag{166}$$

Now, introduce the convolution operator $\mathcal{S}_h$ with kernel $S_h(z) = h^{-1}\,S(h^{-1}z)$, defined via its Fourier transform as

$$\hat{S}(\omega) = \int_{\mathbb{R}^d} S(z)\,\mathrm{e}^{-2\pi i\,\langle z,\omega\rangle}\,d\mu(z) = \left\{1 + |\,2\pi\omega\,|^2\right\}^{-1/2}, \tag{167}$$

for $\omega \in \mathbb{R}^d$. Here and below, $|\,\omega\,|$ denotes the Euclidean norm of ω, and $\langle\,\omega\,,\,z\,\rangle$ denotes the inner product on $\mathbb{R}^d$. In fact, then

$$S(z) = 2^{-(d-1)/2}\,\pi^{-(d+1)/2}\,|\,z\,|^{-(d-1)/2}\,K_{(d-1)/2}(|\,z\,|), \quad z \in \mathbb{R}^d, \tag{168}$$

where K_ν is the modified Bessel function of the second kind of order ν. Aronszajn and Smith [1] turn out to be the ideal reference for this.

The convolution operator is defined as

$$[\mathcal{S}_h f](z) = \int_{\mathbb{R}^d} S_h(z-s)\,f(s)\,d\mu(s), \quad z \in \mathbb{R}^d, \tag{169}$$

and satisfies $(S_h f)\hat{}\,(\omega) = \left\{1 + (2\pi h\,|\,\omega\,|)^2\right\}^{-1/2}\,\hat{f}(\omega)$ for $\omega \in \mathbb{R}^d$.

The net effect is that $v = \mathcal{S}_h u$ satisfies

$$\|\,u\,\|^2 + h^2\,\|\,\nabla u\,\|^2 = \|\,v\,\|^2, \tag{170}$$

so that the final change of variable $f = \mathfrak{M}(w)$, where

$$[\mathfrak{M}(w)](y) = \left\{\left[\mathcal{S}_h\sqrt{w}\,\right](y)\right\}^2, \quad y \in \mathbb{R}^d, \tag{171}$$

transforms the original maximum likelihood problem (161) into

$$\begin{aligned} &\text{minimize} \quad -\frac{1}{n}\sum_{i=1}^{n} \log[\mathcal{K}\mathfrak{M}(w)](Y_i) + \int_{\mathbb{R}^d} w(y)\, d\mu(y) \\ &\text{subject to} \quad w \in \mathcal{P}. \end{aligned} \tag{172}$$

(Actually, the pdf constraints are treated a bit cavalierly. Obviously f and w cannot both be pdfs, but let it pass.)

It now turns out that there is an EM algorithm for the smoothed maximum likelihood problem (172) to wit

$$\begin{aligned} w_{k+1/3} &= \mathfrak{M}(w_k), \\ w_{k+2/3}(z) &= \left\{w_{k+1/3}(y)\right\}^{1/2} \cdot \frac{1}{n}\sum_{i=1}^{n} \frac{k(Y_i - z)}{[\mathcal{K}w_{k+1/3}](Y_j)}, \\ w_{k+1}(z) &= \left\{w_k(z)\right\}^{1/2} [\mathcal{S}_h w_{k+2/3}](z). \end{aligned} \tag{173}$$

It has the same monotonicity properties as the NEMS algorithm; see section "Monotonicity of the Smoothed EM Algorithm." The original method of Miller and Roysam [73] satisfies similar monotonicity properties (assuming that (165) is solved exactly). See section "Monotonicity for Exact Gibbs Smoothing."

Gibbs Smoothing

Whereas Good's roughness penalization was essentially aimed at the continuous setting, attention now turns to a purely discrete point of view. So, let us consider the discrete maximum penalized likelihood problem

$$\begin{aligned} &\text{minimize} \quad -\sum_{j=1}^{m} b_j \, \log[Ax]_j + \sum_{p=1}^{\ell} x_p \; + \; \lambda\, G(x) \\ &\text{subject to} \quad x \in V_\ell, \end{aligned} \tag{174}$$

with V_ℓ given by (43) and $A \in \mathbb{R}^{m\times\ell}$ has components $a(j,p)$ given by (127) and $\lambda > 0$ is the regularization parameter. The typical form of the penalization associated with the name of Gibbs smoothing is

$$G(x) = \sum_{p.q} w_{pq}\, \phi\left(\sigma^{-1}(x_p - x_q)\right), \tag{175}$$

for some convex function ϕ and nonnegative weights w_{pq} and positive σ. Some typical examples for ϕ are $\phi(t) = \log\cosh(t)$ and $\phi(t) = |t|$ for $t \in \mathbb{R}$. The nonzero weights w_{pq} determine a neighborhood system. The neighborhood of the p-

th component of x is given by $\{q \mid w_{pq} > 0\}$. Although x was encoded as a column vector, one should think of x as a two-dimensional image or three-dimensional structure, so that neighboring image elements may have widely differing indices p and q. See [45, 46, 66]. The approach of (174) originated with Green [49].

The role of the penalty term is to penalize differences in neighboring components of x, but large differences are not penalized much more. In fact, this is an argument for choosing $\phi(t) = \min(|t|, \delta)$ for some δ.

To solve the problem (174), again proceed iteratively, and perform the E-step of the EM algorithm. As before, this gives

$$\begin{aligned} &\text{minimize} \quad -\sum_{p=1}^{\ell} \tilde{x}_p^{[k]} \log x_p + \sum_{p=1}^{\ell} x_p + \lambda\, G(x) \\ &\text{subject to} \quad x \in V_\ell, \end{aligned} \tag{176}$$

with $\tilde{x}^{[2]}$ given by (129). For convenience, the constraint that $x \in V_\ell$ is now dropped. To solve the resulting problem, set the gradient equal to 0,

$$-\frac{\tilde{x}_p^{[2]}}{x_p} + 1 + \lambda\, \nabla G(x) = 0. \tag{177}$$

Now, the one-step-late idea of Green [49] is to approximately solve this equation by

$$x_p^{[2]} = \frac{\tilde{x}_p^{[2]}}{1 + \lambda\, [\nabla G(x^{[1]})]_p}, \quad p = 1, 2, \ldots, \ell. \tag{178}$$

This is referred to as OSL-EM. Green [49] reports that this works well for small λ. Regarding its convergence under appropriate conditions, see [62]. If (177) is solved exactly, then the resulting algorithm has the usual nice monotonicity properties; see section "Monotonicity for Exact Gibbs Smoothing."

Hebert and Leahy [54] observed that (178) is similar in spirit to Jacobi's method for solving systems of linear equations, and they noticed that the Gauss–Seidel analogue of sequentially solving (177) speeds up the computations. See also [41]. For other ways to accelerate EM algorithms, see Sect. 10.

8 Convergence of EM Algorithms

The convergence of the Shepp–Vardi EM algorithm is based on two rather remarkable monotonicity properties of the EM algorithm, established using analytical methods by Mülthei and Schorr [75]. Unfortunately, the geometric approach of Csiszár and Tusnády [23] that seems to explain why the Mülthei–Schorr approach works is not discussed. See [38]. However, the methods generalize in different ways. See Sect. 9.

The Two Monotonicity Properties

Consider the discretized maximum likelihood problem of positron emission tomography, repeated here for convenience:

$$\begin{aligned} \text{minimize} \quad & L(x) \stackrel{\text{def}}{=} -\sum_{j=1}^{m} b_j \log[Ax]_j + \sum_{p=1}^{\ell} x_p \\ \text{subject to} \quad & x \in V_\ell, \end{aligned} \tag{179}$$

where $b_j = N_j/N$. Here, V_ℓ is given by (43) and $A \in \mathbb{R}^{m\times\ell}$ has nonnegative components $a(j, p)$ given by (126), with unit column sums

$$\sum_{j=1}^{m} a(j, p) = 1, \quad p = 1, 2, \ldots, \ell. \tag{180}$$

It is clear that the problem (179) is convex and that solutions exist. The uniqueness is guaranteed only if A has full column rank. Regardless, the set of minimizers, denoted by $\mathcal{C}$, is convex.

Recall that the EM algorithm for solving (179) is, for $k = 1, 2, \ldots$,

$$\begin{aligned} x_p^{[k+1]} &= x_p^{[k]} \cdot [A^T r^{[k]}]_p, \quad p = 1, 2, \ldots, \ell, \\ r_p^{[k]} &= \frac{b_j}{[Ax^{[k]}]_j}, \quad j = 1, 2, \ldots, m. \end{aligned} \tag{181}$$

starting from some initial strictly positive probability vector $x^{[1]}$.

The two monotonicity properties are as follows:

$$L(x^{[k]}) - L(x^{[k+1]}) \geqslant \mathrm{KL}(x^{[k+1]}, x^{[k]}) \geqslant 0, \tag{182}$$

and, if x^* is any solution of (179),

$$\mathrm{KL}(x^*, x^{[k]}) - \mathrm{KL}(x^*, x^{[k+1]}) \geqslant L(x^{[k]}) - L(x^*) \geqslant 0. \tag{183}$$

The meaning of the first monotonicity property is clear: It says that the likelihood decreases if successive iterates are different. The second one says that the iterates get closer to every minimizer as measured by the Kullback–Leibler "distance." The everyday image is that if one thinks of the set of minimizers as an airport, then the iterates land like a helicopter, not like an airplane. This kind of monotonicity is called Féjèr monotonicity.

The two monotonicity properties imply that the EM algorithm converges.

Theorem 1. *If $x^{[1]}$ is strictly positive, then the sequence $\{x^{[k]}\}_k$ generated by the* EM *algorithm* (181) *converges to a solution, say x^{**}, of the maximum likelihood problem* (179). *In particular,*

$$\lim_{k\to\infty} \mathrm{KL}(x^{**}, x^{[k]}) = 0.$$

Proof. The first inequality says that the negative log-likelihood is strictly decreasing, unless $x^{[k]} = x^{[k+1]}$. If $x^{[k]} = x^{[k+1]}$ does indeed hold, then the second inequality says that $L(x^{[k]}) = L(x^*)$, so that $x^{[k]}$ is a solution of (179). In general, the second inequality implies that $\{\mathrm{KL}(x^*, x^{[k]})\}_k$ is a decreasing sequence. Since the sequence is bounded from below (by 0), it must have a limit, but then $\mathrm{KL}(x^*, x^{[k]}) - \mathrm{KL}(x^*, x^{[k+1]}) \longrightarrow 0$, which implies that

$$L(x^{[k]}) \longrightarrow L(x^*). \tag{184}$$

So, the negative log-likelihood converges. Finally, since $\sum x_p^{[k]} = 1$, the sequence $\{x^{[k]}\}_p$ is bounded and hence has a convergent subsequence, say, with limit x^{**}. By (184), then $L(x^{**}) = L(x^*)$, so that x^{**} is a minimizer also. Now, in the second monotonicity property, one may replace x^* by x^{**}, and then $\{\mathrm{KL}(x^{**}, x^{[k]})\}_k$ is decreasing. Since a subsequence converges to 0, then the whole sequence converges to 0. ■

It should be observed that the theorem is actually not very useful: When using the algorithm (181), one will *always* stop the algorithm well short of convergence. See, e.g., [69, 80]. Thus, the existence of maximum likelihood estimators is moot. One may think of this as an unfortunate side effect of discretization. For the continuous version, say, for the deconvolution problem, one does indeed have the analogues of the above two monotonicity properties, but the second one is vacuous, since the continuous maximum likelihood problem has no solutions. For the NEMS algorithm, one can show the existence of solutions as well as its convergence by way of the two monotonicity properties. See section "Monotonicity of the Smoothed EM Algorithm."

In the following subsections, the two monotonicity properties are proved for the standard discrete Shepp–Vardi EM algorithm, for the continuous version of the NEMS algorithm, and for the exact version of Gibbs smoothing (but not for the one-step-late version). The basic tool is the analytical proof of Mülthei and Schorr [75], which is actually quite versatile, as demonstrated in Sect. 9.

Monotonicity of the Shepp–Vardi EM Algorithm

Here, the two monotonicity properties of the EM algorithm are exhibited, following the proof of Mülthei and Schorr [75]. The first monotonicity property (182) follows from the derivation of the E-step of the EM algorithm. However, here a purely

analytical proof is explained. Vardi et al. [96] prove the two monotonicity properties using the geometric results of Csiszár and Tusnády [23].

It is useful to define the operator R on nonnegative vectors $x \in \mathbb{R}^{\ell}$ by

$$[R\,x]_p = x_p \left[A^T\,(b/Ax)\right]_p, \quad p = 1, 2, \ldots, \ell, \tag{185}$$

where b/Ax denotes the vector of componentwise quotients.

Lemma 1. *For all nonnegative x and y, with y strictly positive,*

$$L(x) \leqslant L(y) + \mathrm{KL}(R\,y, x) - \mathrm{KL}(R\,y, y).$$

Proof. Note that for all nonnegative vectors x and y,

$$L(x) - L(y) = -\sum_{j=1}^{m} b_j\,\log\frac{[Ax]_j}{[Ay]_j} + \sum_{p=1}^{\ell} x_p - y_p.$$

Now, for strictly positive y, one may write

$$\frac{[Ax]_j}{[Ay]_j} = \sum_{p=1}^{\ell} \frac{a(j,p)\,y_p}{[Ay]_j} \cdot \frac{x_p}{y_p}.$$

For each j, this is a convex combination of the points x_p/y_p, $p = 1, 2, \ldots, \ell$. Since $t \longmapsto -\log t$ is convex, then by Jensen's inequality

$$\begin{aligned} L(x) - L(y) &\leqslant -\sum_{j=1}^{m} b_j \sum_{p=1}^{\ell} \frac{a(j,p)\,y_p}{[Ay]_j}\,\log\frac{x_p}{y_p} + \sum_{p=1}^{\ell} x_p - y_p \\ &\leqslant \sum_{p=1}^{\ell} -y_p\,[A^T r]_p\,\log\frac{x_p}{y_p} + x_p - y_p, \end{aligned}$$

where in the last step the order of summation was interchanged. The lemma follows. ∎

Proof of the first monotonicity property (182). In the inequality of the lemma above, take $y = x^{[k]}$ and $x = R\,y = R\,x^{[k]} = x^{[k+1]}$. Then, $L(x^{[k+1]}) - L(x^{[k]}) \leqslant -\mathrm{KL}(x^{[k+1]}, x^{[k]})$. ∎

Proof of the second monotonicity property (183). Start with

$$\mathrm{KL}(x^*, x^{[k]}) - \mathrm{KL}(x^*, x^{[k+1]}) = \sum_{p=1}^{\ell} x_p^* \log \frac{x_p^{[k+1]}}{x_p^{[k]}}$$
$$= \sum_{p=1}^{\ell} x_p^* \log \left[A^T \left\{ \frac{b}{Ax^{[k]}} \right\} \right]_p .$$

Now, if x^* solves (179), then it must satisfy the necessary and sufficient conditions for a minimum

$$x^* \geqslant 0, \quad \nabla L(x^*) \geqslant 0, \quad x_p^* [\nabla L(x^*)]_p = 0 \quad \text{for all} \quad p.$$

The last condition says that $x_p^* \left(-[A^T r^*]_p + 1 \right) = 0$, where $r_j^* = b_j / [Ax^*]_j$ for all j, so that if $x_p^* > 0$, then $[A^T r^*]_p = 1$. So, for $x_p^* > 0$, write

$$\left[A^T \left\{ \frac{b}{x^{[k]}} \right\} \right]_p = \sum_{j=1}^{m} \frac{a(j,p)\, b_j}{[Ax^*]_j} \cdot \frac{[Ax^*]_j}{[Ax^{[k]}]_j},$$

which is a convex combination of the points $[Ax^*]_j / [Ax^{[k]}]_j$, so by the concavity of the logarithm,

$$\sum_{p=1}^{\ell} x_p^* \log \left[A^T \left\{ \frac{b}{Ax^{[k]}} \right\} \right]_p \geqslant \sum_{p=1}^{\ell} x_p^* \sum_{j=1}^{m} \frac{a(j,p)\, b_j}{[Ax^*]_j} \cdot \log \frac{[Ax^*]_j}{[Ax^{[k]}]_j}$$
$$\geqslant \sum_{j=1}^{m} b_j \log \frac{[Ax^*]_j}{[Ax^{[k]}]_j} = \mathrm{KL}(b, Ax^{[k]}) - \mathrm{KL}(b, Ax^{[k+1]}),$$

where the last equality follows from $\sum x_p^{[k]} = \sum x_p^{[k+1]} = \sum b_j$. ■

Monotonicity for Mixtures

Here, the two monotonicity properties of the EM algorithm for mixtures of known densities are discussed. The difference with the Shepp–Vardi EM algorithm is that the system matrix is not normalized to have unit column sums. It will transpire that this does not make any difference.

Recall that the problem is to estimate the pdf

$$f_Y(y) = \sum_{j=1}^{m} x_{o,j}\, a_j(y), \quad y \in \mathbb{R}^d, \tag{186}$$

where the a_j are known pdfs and x_o is an unknown probability vector, given a random sample $Y_1, Y_2, \ldots, Y_n$ of the random variable Y with density f_Y. Define the matrix $A \in \mathbb{R}^{n \times m}$ by

$$A_{ij} = a_{ij} = a_j(Y_i) \quad \text{for all } i \text{ and } j. \tag{187}$$

The EM algorithm for estimating x_o is, starting from the uniform vector $x^{[1]}$,

$$x^{[k+1]} = M x^{[k]}, \tag{188}$$

where the iteration operator M is defined as

$$[Mx]_j = x_j \cdot \frac{1}{n} \sum_{i=1}^{n} \frac{a_{ij}}{[Ax]_i}, \quad j = 1, 2, \ldots, m. \tag{189}$$

One begins again with deriving the majorizing function inequality. However, first replace the maximum likelihood problem (44) by the equivalent

$$\begin{aligned} &\text{minimize} \quad L_n(x) \stackrel{\text{def}}{=} -\frac{1}{n} \sum_{i=1}^{n} \log \left(\sum_{j=1}^{m} x_j \, a_{ij} \right) + \sum_{j=1}^{m} x_j \\ &\text{subject to} \quad x \geqslant 0. \end{aligned} \tag{190}$$

Note that the constraint that x be a probability vector was traded for the added sum in the objective function.

The majorizing function inequality is the same as before, as is its proof. Then, the first monotonicity property follows.

Lemma 2. *If x and y are nonnegative probability vectors, with y strictly positive, then*

$$L_n(x) \leqslant L_n(y) + \mathrm{KL}(M\, y, x) - \mathrm{KL}(M\, y, y).$$

Note that the minimizer of the right-hand side (over x) is $x = My$.

Lemma 3. *Starting from a strictly positive $x^{[1]}$, the iterates of the* EM *algorithm* (188) *satisfy*

$$L_n(x^{[k]}) - L_n(x^{[k+1]}) \geqslant \mathrm{KL}(x^{[k+1]}, x^{[k]}) \geqslant 0.$$

The second monotonicity property is the same also, but there is a slight change in its proof.

Lemma 4. *Let* x^* *be a solution of* (190). *Starting from a strictly positive* $x^{[1]}$, *the iterates of the* EM *algorithm* (188) *satisfy*

$$\mathrm{KL}(x^*, x^{[k]}) - \mathrm{KL}(x^*, x^{[k+1]}) \geqslant L_n(x^{[k]}) - L_n(x^*) \geqslant 0.$$

Proof. Since $\sum x_j^{[k]} = \sum x_j^{[k+1]} = 1$, one has as usual

$$\begin{aligned}\mathrm{KL}(x^*, x^{[k]}) - \mathrm{KL}(x^*, x^{[k+1]}) &= \sum_{j=1}^{m} x_p^* \log \frac{x_p^{[k+1]}}{x_p^{[k]}} \\ &= \sum_{j=1}^{\ell} x_j^* \log \left[A^T \left\{ \frac{(1/n)}{Ax^{[k]}} \right\} \right]_j .\end{aligned}$$

Now, if x^* solves (179), then it must satisfy the necessary and sufficient conditions for a minimum

$$x^* \geqslant 0, \quad \nabla L_n(x^*) \geqslant 0, \quad x_j^* \, [\nabla L_n(x^*)]_j = 0 \quad \text{for all} \quad j.$$

The last condition says that $x_j^* \left(-[A^T r^*]_j + 1\right) = 0$, where $r_i^* = (1/n)/[Ax^*]_i$ for all i, so that if $x_j^* > 0$, then $[A^T r^*]_j = 1$. So, for $x_j^* > 0$, write

$$\left[A^T \left\{ \frac{(1/n)}{Ax^{[k]}} \right\} \right]_j = \sum_{i=1}^{n} \frac{(1/n)\, a_{ij}}{[Ax^*]_i} \cdot \frac{[Ax^*]_i}{[Ax^{[k]}]_i},$$

which is a convex combination of the points $[Ax^*]_i / [Ax^{[k]}]_i$, so by the concavity of the logarithm,

$$\begin{aligned}\sum_{j=1}^{m} x_j^* \log \left[A^T \left\{ \frac{(1/n)}{Ax^{[k]}} \right\} \right]_j &\geqslant \sum_{j=1}^{m} x_j^* \sum_{i=1}^{n} \frac{(1/n)\, a_{ij}}{[Ax^*]_i} \cdot \log \frac{[Ax^*]_i}{[Ax^{[k]}]_i} \\ &\geqslant \sum_{i=1}^{n} (1/n) \log \frac{[Ax^*]_i}{[Ax^{[k]}]_i} + \sum_{j=1}^{m} x_j^{[k]} - x_j^* = L_n(x^{[k]}) - L_n(x^*),\end{aligned}$$

where the last equality follows from $\sum x_j^{[k]} = \sum x_j^* = 1$. ∎

The convergence of the iterates of the EM algorithm follows.

Monotonicity of the Smoothed EM Algorithm

Here, the monotonicity properties of the NEMS algorithm for the smoothed maximum likelihood problem (158) are proved. The problem is

$$\begin{aligned} &\text{minimize} \quad L_n(f) \stackrel{\text{def}}{=} -\frac{1}{n}\sum_{i=1}^{n} \log[\mathcal{K}\mathcal{N}(f)](W_i) + \int_{\mathbb{R}^d} f(y)\, d\mu(y) \\ &\text{subject to} \quad f \in L^1(\mathbb{R}^d), \quad f \geqslant 0. \end{aligned} \tag{191}$$

Since this is an infinite-dimensional problem, showing that the iterates of the NEMS algorithm converge to a solution of (191) is a bit more involved. In particular, it requires us to show the existence of solutions. The only remarkable thing about the proofs of the two monotonicity properties for the NEMS algorithm is that apart from a few cosmetic changes, they are exactly the same as for the Shepp–Vardi EM algorithm. The argument follows Eggermont [34].

The NEMS algorithm (159) may be represented as

$$g_{k+1} = \mathcal{T}_h f_k, \quad f_{k+1} = \mathcal{S}_h\, g_{k+1}, \tag{192}$$

starting from a strictly positive initial guess f_1, assumed to be a pdf. Here, the map $\mathcal{T}_h$ is defined as

$$[\mathcal{T}_h f](z) = [\mathcal{N}(f)](z) \cdot \frac{1}{n}\sum_{i=1}^{n} \frac{k(W_i - z)}{[\mathcal{K}\mathcal{N}(f)](W_i)}. \tag{193}$$

The claim is now that the iterates of the NEMS algorithm satisfy the same two monotonicity properties (182) and (183). The crux is again an analytical proof of what amounts to the E-step of the EM algorithm.

Lemma 5. *For all densities φ and ψ,*

$$L_n(\varphi) \leqslant L_n(\psi) + \mathrm{KL}(\mathcal{S}_h\mathcal{T}_h\psi, \varphi) - \mathrm{KL}(\mathcal{S}_h\mathcal{T}_h\psi, \psi).$$

Proof. Similar to the proof of Lemma 1, one gets that

$$L_n(\varphi) - L_n(\psi) \leqslant -\int_{\mathbb{R}^d} [\mathcal{T}_h\psi](z)\, \log \frac{[\mathcal{N}(\varphi)](z)}{[\mathcal{N}(\psi)](z)}\, d\mu(z).$$

Since

$$\log\left([\mathcal{N}(\varphi)](z)/[\mathcal{N}(\psi)](z)\right) = [\mathcal{S}_h \log(\varphi/\psi)](z),$$

and $\mathcal{S}_h$ is a symmetric operator, then

$$-\int_{\mathbb{R}^d} [\mathcal{T}_h\psi](z)\, \log \frac{[\mathcal{N}(\varphi)](z)}{[\mathcal{N}(\psi)](z)}\, d\mu(z) = -\int_{\mathbb{R}^d} [\mathcal{S}_h\mathcal{T}_h\psi](y)\, \log \frac{\varphi(y)}{\psi(y)}\, d\mu(y),$$

and the lemma follows. ■

Lemma 6. *The iterates f_k generated by the* NEMS *algorithm* (192) *satisfy*

$$L_n(f_k) - L_n(f_{k+1}) \geqslant \mathrm{KL}(f_{k+1}, f_k) \geqslant 0.$$

Proof. In Lemma 5, take $\varphi = f_{k+1}$ and $\psi = f_k$. Then, $\mathcal{S}_h\mathcal{T}_h\psi = f_{k+1}$. ■

The second monotonicity property is actually a little bit stronger than the one for the Shepp–Vardi EM algorithm. Note that by the convexity of the KL function jointly in both its arguments, $\mathrm{KL}(\mathcal{S}_h\varphi, \mathcal{S}_h\psi) \leqslant \mathrm{KL}(\varphi, \psi)$.

Lemma 7. *Let f^* be a solution of* (191)*, with $\mathrm{KL}(f^*, f_1) < \infty$. Then, the iterates f_k generated by the* NEMS *algorithm* (192) *satisfy*

$$\begin{aligned}\mathrm{KL}(f^*, f_k) - \mathrm{KL}(f^*, f_{k+1}) &\geqslant \mathrm{KL}(f^*, f_k) - \mathrm{KL}(\mathcal{T}_h f^*, \mathcal{T}_h f_k)\\ &\geqslant L_n(f_k) - L_n(f^*) \geqslant 0.\end{aligned}$$

Proof. Start in the usual fashion and obtain

$$\begin{aligned}L_n(f_k) - L_n(f^*) &= \frac{1}{n}\sum_{i=1}^{n} \log \frac{[\mathcal{K}\mathcal{N} f^*](W_i)}{[\mathcal{K}\mathcal{N} f_*](W_i)}\\ &= \frac{1}{n}\sum_{i=1}^{n} \frac{[\mathcal{K}\mathcal{N} f^*](W_i)}{[\mathcal{K}\mathcal{N} f^*](W_i)} \log \frac{[\mathcal{K}\mathcal{N} f^*](W_i)}{[\mathcal{K}\mathcal{N} f_k](W_i)}\\ &= \int_{\mathbb{R}^d} [\mathcal{N} f^*](z) \cdot \frac{1}{n}\sum_{i=1}^{n} \frac{k(W_i - z)}{[\mathcal{K}\mathcal{N} f^*](z)} \log \frac{[\mathcal{K}\mathcal{N} f^*](W_i)}{[\mathcal{K}\mathcal{N} f_k](W_i)}\, d\mu(z).\end{aligned}$$

Now, one would like to get a convex combination, so multiply and divide by the sum of the weights $k(W_i - z)/[\mathcal{K}\mathcal{N} f^*](W_i)$. Then, the concavity of the logarithm gives that the last expression is dominated by

$$\int_{\mathbb{R}^d} [\mathcal{N} f^*](z) \cdot \frac{1}{n}\sum_{i=1}^{n} \frac{k(W_i - z)}{[\mathcal{K}\mathcal{N} f^*](z)} \log \left(\frac{\dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \dfrac{k(W_i - z)}{[\mathcal{K}\mathcal{N} f_k](z)}}{\dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \dfrac{k(W_i - z)}{[\mathcal{K}\mathcal{N} f^*](z)}} \right) d\mu(z).$$

Now, this expression may be cleaned up as

$$\int_{\mathbb{R}^d} [\mathcal{T}_h f^*](z) \log \left(\frac{[\mathcal{T}_h f_k](z)}{[\mathcal{N} f_k](z)} \cdot \frac{[\mathcal{N} f^*](z)}{[\mathcal{T}_h f^*](z)} \right) d\mu(z).$$

After splitting up the logarithm, note that

$$\int_{\mathbb{R}^d} [\mathcal{T}_h f^*](z) \log\left(\frac{[\mathcal{T}_h f_k](z)}{[\mathcal{T}_h f^*](z)}\right) d\mu(z) = -\mathrm{KL}(\mathcal{T}_h f^*, \mathcal{T}_h f_k)$$

because $\mathcal{T}_h \varphi$ is a pdf if φ is one and also

$$\int_{\mathbb{R}^d} [\mathcal{T}_h f^*](z) \log\left(\frac{[\mathcal{N} f^*](z)}{[\mathcal{N} f_k](z)}\right) d\mu(z) = \int_{\mathbb{R}^d} [\mathcal{T}_h g^*](z) \left[\mathcal{S}_h \log\frac{f^*}{f_k}\right](z)\, d\mu(z)$$
$$= \int_{\mathbb{R}^d} [\mathcal{S}_h \mathcal{T}_h f^*](y) \log\frac{f^*(y)}{f_k(y)}\, d\mu(y) = \mathrm{KL}(f^*, f_k),$$

since $\mathcal{S}_h \mathcal{T}_h f^* = f^*$. Putting all of this together shows that

$$L_n(f_k) - L_n(f^*) \leqslant \mathrm{KL}(f^*, f_k) - \mathrm{KL}(\mathcal{T}_h f^*, \mathcal{T}_h f_k),$$

and the lemma follows. ■

The two monotonicity properties imply that the NEMS algorithm converges to a solution of the smoothed maximum likelihood problem (191) and that this problem actually has a solution.

Theorem 2. *The smoothed maximum likelihood problem* (191) *has a solution* $f^* \in L^1(\mathbb{R}^d)$.

Proof. One first shows that $L_n(f)$ is bounded from below. Let f be a pdf on $\mathbb{R}^d$ such that $[\mathcal{K}\mathcal{N} f](W_i) > 0$ for all i. Let $\mathbf{1} \in \mathbb{R}^n$ be the vector of all ones, and let $v_i = [\mathcal{K}\mathcal{N} f](W_i)$. Then,

$$L_n(f) = \mathrm{KL}\left(\frac{1}{n}\mathbf{1}, v\right) - \frac{1}{n}\sum_{i=1}^n \log[\mathcal{K}\mathcal{N} f](W_i) + \int_{\mathbb{R}^d} f(y)\, d\mu(y).$$

Now, by convexity $[\mathcal{N} f](z) \leqslant [\mathcal{S}_h f](z)$ for all z, so that

$$\begin{aligned}
[\mathcal{K}\mathcal{N} f](z) &= \int_{\mathbb{R}^d} k(W_i - z)\, [\mathcal{N} f](z)\, d\mu(z) \\
&\leqslant \int_{\mathbb{R}^d} k(W_i - z)\, [\mathcal{S}_h f](z)\, d\mu(z) \\
&= \int_{\mathbb{R}^d} f(y) \int_{\mathbb{R}^d} k(W_i - z)\, S_h(z - y)\, d\mu(y)\, d\mu(z) \\
&\leqslant \mu_h \int_{\mathbb{R}^d} f(y)\, d\mu(y) = \mu_h,
\end{aligned}$$

where

$$\mu_h = \sup_{y\in\mathbb{R}^d} \int_{\mathbb{R}^d} k(W_i - z)\, S_h(z-y)\, d\mu(y)\, d\mu(z) \leqslant \sup_{y\in\mathbb{R}^d} S_h(z),$$

since k is a pdf. Since $S_h(z) = h^{-d} S(h^{-1}z)$, the boundedness of S then gives that $\mu_h < \infty$ for fixed $h > 0$. It follows that $L_n(f)$ is bounded from below.

Now, let $\{\varphi_k\}_k$ be a minimizing sequence for $L_n(f)$. Apply one step of the NEMS algorithm to each φ_k, so $\psi_k = \mathcal{S}_h\mathcal{T}_h\varphi_k$, $k = 1, 2, \ldots$. By the first monotonicity property, then $\{\psi_k\}_k$ is a minimizing sequence also. Since each $\mathcal{T}_h\varphi_k$ is a pdf, then the ψ_k are uniformly continuous on $\mathbb{R}^d$, and so it has a subsequence which converges in the strict topology, i.e., uniformly on every compact subset of $\mathbb{R}^d$, say, with limit ψ^*. This is the Arzelà–Ascoli theorem for the strict topology; see [2]. Then, along this subsequence

$$[\mathcal{KN}(\psi_k)](W_i) \longrightarrow [\mathcal{KN}(\psi^*)](W_i),$$

and it follows that again along this same subsequence

$$L_n(\psi_k) \longrightarrow L_n(\psi^*).$$

Since the whole sequence $\{\psi_k\}_k$ was a minimizing sequence, this shows that ψ^* solves the problem (191). ■

Theorem 3. *For $f^{[1]}$ strictly positive with $\mathrm{KL}(f^*, f^{[1]}) < \infty$, the* NEMS *algorithm converges to a solution of* (191).

Proof. The proof is just about the same as for the discrete EM algorithm. Thus, the first monotonicity property shows that $\{L_n(f_k)\}$ is decreasing. The second monotonicity property shows that $\{\mathrm{KL}(f^*, f_k)\}_k$ is decreasing as well and so has a nonnegative limit. But then $\mathrm{KL}(f^*, f_k) - \mathrm{KL}(f^*, f_{k+1})$ converges to 0, so that again the second monotonicity property gives that the NEMS sequence $\{f_k\}_k$ is a minimizing sequence. All one has to do is extract a convergent subsequence, but that follows from the argument in the proof of the existence of convergent subsequences. Thus, there exists a subsequence which converges to some element f^{**} in the strict topology. Then, $[\mathcal{KN}(f_k)](W_i) \longrightarrow [\mathcal{KN}(f^{**})](W_i)$ for all i, and then $L_n(f_k) \longrightarrow L_n(f^{**})$ initially only along the subsequence, but since $\{L_n(f_k)\}_k$ is decreasing, then along the whole sequence. Now, if $\mathrm{KL}(f^{**}, f_1) < \infty$, then apply the second monotonicity property with the solution f^{**}, and then one finds that $\mathrm{KL}(f^{**}, f_k) \longrightarrow 0$ along the subsequence, but since $\{\mathrm{KL}(f^{**}, f_k)\}_k$ is decreasing, then along the whole sequence.

If $\mathrm{KL}(f^{**}, f_1) = \infty$, then for $0 < \varepsilon < 1$ but arbitrary, apply the second monotonicity property with the solution $f_\varepsilon^* = \varepsilon f^* + (1-\varepsilon) f^{**}$. Then, $\mathrm{KL}\left(f_\varepsilon^*, f_1\right) < \infty$ and $\mathrm{KL}\left(f_\varepsilon^*, f_k\right)$ converges, and it is easy to see that

$$\lim_{k\to\infty} \mathrm{KL}\left(f_\varepsilon^*, f_k\right) = \mathrm{KL}\left(f_\varepsilon^*, f^{**}\right) = o(1) \quad \text{for} \quad \varepsilon \longrightarrow 0.$$

It follows that $\mathrm{KL}(f^{**}, f_k) \longrightarrow 0$. ■

Monotonicity for Exact Gibbs Smoothing

The two monotonicity properties also hold for penalized maximum likelihood estimation with Gibbs smoothing, at least if the M-step of the EM algorithm is performed exactly; see (196) below. This is at least approximately the case in the approach of Miller and Roysam [73] but not so for the one-step-late approach of Green [49].

Here, the monotonicity properties are proved for "arbitrary" Gibbs functionals. So, consider the maximum penalized likelihood problem for emission tomography

$$\begin{aligned} &\text{minimize} \quad \Lambda(x) \stackrel{\text{def}}{=} -\sum_{j=1}^{m} b_j \log[Ax]_j + \sum_{p=1}^{\ell} x_p + G(x) \\ &\text{subject to} \quad x \geqslant 0, \end{aligned} \tag{194}$$

where $G(x)$ is convex and differentiable and satisfies

$$\lim_{\|x\|\to\infty} G(x) = +\infty. \tag{195}$$

Assume that b is strictly positive and that A satisfies the usual conditions (126). Note that typically, the roughness prior will be of the form $\lambda G(x)$ for some small positive parameter λ. In the present context, one may as well take $\lambda = 1$.

The goal is again to derive the two monotonicity properties. The majorizing functional inequality is just about the same as for the Shepp–Vardi EM algorithm; see (180). Recall the definition of the operator R from (185).

Lemma 8. *For all probability vectors x and y,*

$$\Lambda(x) \leqslant \Lambda(y) + \mathrm{KL}(Ry, x) - \mathrm{KL}(Ry, y) + G(x) - G(y),$$

where $r_j = b_j/[Ay]_j$ for $j = 1, 2, \ldots, m$.

Now, to minimize the right-hand side, set the gradient with respect to x equal to 0. With $y = x^{[k]}$, this gives the next iterate implicitly as

$$x_p^{[k+1]} = \frac{x_p^{[k]}\,[A^T r^{[k]}]_p}{1 + [\nabla G(x^{[k+1]})]_p}, \quad p = 1, 2, \ldots, \ell, \tag{196}$$

with $r_j^{[k]} = b_j/[Ax^{[k]}]_j$ for all j. Note that $1 + \left[\nabla G\left(x^{[k+1]}\right)\right]_p > 0$ for all p, since the equations

$$x_p^{[k+1]}\left(1 + \left[\nabla G\left(x^{[k+1]}\right)\right]_p\right) = x_p^{[k]}, \quad p = 1, 2, \ldots, \ell,$$

have a solution (the minimization problem has a solution) and $x^{[k]}$ is strictly positive (by induction).

The first monotonicity property is almost immediate. The second one takes more work. For nonnegative vectors x, y, and w, define

$$\mathrm{KL}(x, y|w) = \sum_{p=1}^{\ell} w_p \left\{x_p \log \frac{x_p}{y_p} + y_p - x_p\right\}. \tag{197}$$

Lemma 9. *Starting with a strictly positive initial vector* $x^{[1]}$,

$$\Lambda(x^{[k]}) - \Lambda(x^{[k+1]}) \geqslant \mathrm{KL}(x^{[k+1]}, x^{[k]} \mid w^{[k+1]}) \geqslant 0,$$

where $w^{[k+1]} = 1 + \nabla G(x^{[k+1]})$.

Proof. From Lemma 8, one gets

$$\begin{aligned}
&\Lambda\left(x^{[k+1]}\right) - \Lambda\left(x^{[k]}\right) \\
&\leqslant \sum_{p=1}^{\ell} \left\{\left(w_p^{[k+1]} x_p^{[k+1]} \log \frac{x_p^{[k]}}{x}^{[k+1]}_p\right) + x_p^{[k+1]} - x_p^{[k]}\right\} \\
&+ G\left(x^{[k+1]}\right) - G\left(x^{[k]}\right).
\end{aligned}$$

Now use that $G(x^{[k+1]}) - G(x^{[k]}) \leqslant \langle \nabla G(x^{[k+1]}), x^{[k+1]} - x^{[k]} \rangle$. ■

Lemma 10. *Let* x^* *be a solution of* (194). *Then, starting from a strictly positive initial guess* $x^{[1]}$,

$$\mathrm{KL}(w^* \cdot x^*, w^{[k]} \cdot x^{[k]}) - \mathrm{KL}(w^* \cdot x^*, w^{[k+1]} \cdot x^{[k+1]}) \geqslant \Lambda(x^{[k]}) - \Lambda(x^*) \geqslant 0,$$

where $w^* = 1 + \nabla G(x^*)$.

Proof. Write $\Lambda(x) = L(x) + G(x)$. So $L(x)$ is the unpenalized negative log-likelihood. Now, as in section "Monotonicity of the Shepp–Vardi EM Algorithm," one has

$$L(x^{[k]}) - L(x^*) = \sum_{j=1}^{m} \left\{ b_j \log \frac{[Ax^*]_j}{[Ax^{[k]}]_j} + [Ax^{[k]}]_j - [Ax^*]_j \right\}$$

$$= \sum_{p=1}^{\ell} \left\{ x_p^* \left[A^T \left\{ r^* \log \frac{Ax^*}{Ax^{[k]}} \right\} \right]_p + x_p^{[k]} - x_p^* \right\},$$

with $r_j^* = b_j / [Ax^*]_j$ for all j. Now, x^* solves the problem (194), so by the previous lemma, it must be a fixed point of the algorithm. So,

$$x_p^* = \frac{x_p^* [A^T r^*]_p}{1 + [\nabla G(x^*)]_p}, \quad p = 1, 2, \ldots, \ell.$$

Then, if $x_p^* > 0$, one must have $[A^T r^*]_p / (1 + [\nabla G(x^*)]_p) = 1$. Consequently, by convexity

$$\frac{\left[A^T \left\{ r^* \log \frac{Ax^*}{Ax^{[k]}} \right\} \right]_p}{1 + [\nabla G(x^*)]_p} \leqslant \log \frac{\left[A^T \left\{ r^* \frac{Ax^*}{Ax^{[k]}} \right\} \right]_p}{1 + [\nabla G(x^*)]_p},$$

which equals

$$\log \left(\frac{x_p^{[k+1]}}{x_p^{[k]}} \cdot \frac{w_p^{[k+1]}}{w_p^*} \right) = \log \left(\frac{w_p^{[k+1]} x_p^{[k+1]}}{w_p^{[k]} x_p^{[k]}} \right) + \log \frac{w_p^{[k]}}{w_p^*}.$$

Now, substitute this in the upper bound for $L(x^{[k]}) - L(x^*)$. This yields

$$L(x^{[k]}) - L(x^*) \leqslant \sum_{p=1}^{\ell} w_p^* x_p^* \log \frac{w_p^{[k+1]} x_p^{[k+1]}}{w_p^{[k]} x_p^{[k]}} +$$

$$\sum_{p=1}^{\ell} \left\{ w_p^* x_p^* \log \frac{w_p^{[k]}}{w_p^*} + x_p^{[k]} - x_p^* \right\}.$$

Now, after some bookkeeping, the first sum is seen to be equal to

$$\mathrm{KL}\left(w^* \cdot x^*, w^{[k]} \cdot x^{[k]} \right) - \mathrm{KL}\left(w^* \cdot x^*, w^{[k+1]} \cdot x^{[k+1]} \right)$$

$$+ \sum_{p=1}^{\ell} \left\{ w_p^{[k+1]} x_p^{[k+1]} - w_p^{[k]} x_p^{[k]} \right\}.$$

Using the inequality $\log t \leqslant t - 1$, one gets that

$$L(x^{[k]}) - L(x^*) \leqslant \mathrm{KL}(w^* \cdot x^*, w^{[k]} \cdot x^{[k]}) - \mathrm{KL}(w^* \cdot x^*, w^{[k+1]} \cdot x^{[k+1]}) + \mathrm{rem} \tag{198}$$

with the remainder

$$\mathrm{rem} = \sum_{p=1}^{\ell} \left\{ (w^{[k]} - 1) \left(x_p^* - x_p^{[k]} \right) - w_p^* x_p^* + w_p^{[k+1]} x_p^{[k+1]} \right\}.$$

Now, since $w_p^{[k+1]} x_p^{[k+1]} = x_p^{[k]} [A^T r^{[k]}]_p$ and likewise for $w_p^* x_p^*$, then

$$\sum_{p=1}^{\ell} w_p^* x_p^* = \sum_{j=1}^{m} b_j = \sum_{p=1}^{\ell} w_p^{[k+1]} x_p^{[k+1]}.$$

The remaining terms add up to $\langle \nabla G(x^{[k]}), x^* - x^{[k]} \rangle$ which is bounded by $G(x^*) - G(x^{[k]})$ (by convexity). Moving this to the left-hand side of the resulting inequality proves the lemma. ■

The convergence of the *exact* EM algorithm with Gibbs smoothing now follows. Lange [62] proves the convergence of the one-step-late version of the algorithm by essentially "soft" methods. It would be nice to see under what conditions the two monotonicity properties carry over to this version.

9 EM-Like Algorithms

The analytical proofs of the inequalities of Lemmas 1 and 5 may be extended to other interesting minimization problems. Rather surprisingly, some of these algorithms enjoy the "same" two monotonicity properties as the EM and NEMS algorithms (and the proofs appear to be simpler). The problems under consideration are "positive" least-squares problems and minimum cross-entropy problems. The main idea is that of majorizing functions, as originally exploited by De Pierro [29] in the maximum penalized likelihood approach for emission tomography.

Again, it would have been nice to also outline the geometric approach of Csiszár and Tusnády [23], which, just like the analytical approach of Mülthei and Schorr [75], is applicable to the minimum cross-entropy problems. However, it is not clear that the Csiszár–Tusnády approach works for the "positive" least-squares problem: The Kullback–Leibler distance shows up in the monotonicity properties. This is in effect due to the multiplicative nature of the algorithms, as explained in the last section.

Minimum Cross-Entropy Problems

Consider again the system of equations

$$Ax = b, \tag{199}$$

in the emission tomography setup (see (126)), with b a nonnegative vector. The interest is in the following minimization problem:

$$\text{minimize} \quad \mathrm{CE}(x) \stackrel{\text{def}}{=} \mathrm{KL}(Ax, b) \quad \text{subject to} \quad x \in \mathbb{R}^{\ell}\,,\ x \geqslant 0. \tag{200}$$

Here, "CE" stands for cross-entropy. (Why it makes sense to consider $\mathrm{KL}(Ax, b)$ instead of $\mathrm{KL}(b, Ax)$ or even $\| Ax - b \|^2$ is not the issue here.)

The objective is to obtain a majorizing function for $\mathrm{CE}(x)$ that would result in a nice algorithm satisfying the "two" monotonicity properties similar to the EM algorithm.

Prejudicing the proceedings somewhat, it is useful to define the operator R on nonnegative vectors by

$$[Ry]_p = y_p \exp\left(\left[A^T \log \frac{b}{Ay}\right]_p\right), \quad p = 1, 2, \ldots, \ell. \tag{201}$$

Here, and elsewhere, $A^T \log(Ay/b) = A^T v$, with $v_j = \log([Ay]_j / b_j)$ for all j.

Lemma 11. *For all nonnegative x, $y \in \mathbb{R}^{\ell}$,*

$$\mathrm{CE}(x) \leqslant \mathrm{CE}(y) + \mathrm{KL}(x, Ry) - \mathrm{KL}(y, Ry).$$

Proof. The starting point is the straightforward identity

$$\mathrm{CE}(x) = \mathrm{CE}(y) + \mathrm{KL}(Ax, Ay) + \langle\, x - y\,,\, A^T r\,\rangle.$$

with $r_j = \log([Ay]_j / b_j)$ for all j.

Now, by convexity of the KL function jointly in both its arguments, the conditions (126) and (127) on A imply that

$$\mathrm{KL}(Ax, Ay) \leqslant \mathrm{KL}(x, y). \tag{202}$$

Finally, since $[A^T r]_p = -\log\left([Ry]_p / y_p\right)$, for all p, then

$$\mathrm{KL}(x, y) + \langle\, x - y\,,\, A^T r\,\rangle = \mathrm{KL}(x, Ry) - \mathrm{KL}(y, Ry).$$

This completes the proof of the lemma. ■

The inequality of the lemma immediately suggests an iterative algorithm for the minimization of $\text{CE}(x)$. Minimizing the right-hand side gives the optimal x as $x = Ry$ with R given by (201). Thus, the iterative algorithm is, starting from a strictly positive $x^{[1]} \in \mathbb{R}^\ell$,

$$x_p^{[k+1]} = [Rx^{[k]}]_p, \quad p = 1, 2, \ldots. \tag{203}$$

This is the simultaneous multiplicative algebraic reconstruction technique (SMART algorithm). It first appeared in [86]; see (also) Holte et al. [57] and Darroch and Ratcliff [25], who called it the iterative rescaling algorithm. The row-action version (MART) originated with Gordon et al. [48]. Byrne [7] developed block-iterative Versions; see section "The MART and SMART Methods." The starting point of Censor and Segman [18] was entropy maximization subject to the linear constraints $Ax = b$ and arrived at various versions of MART including simultaneous and block-iterative versions.

Onto the two monotonicity properties, the first one is immediate.

Lemma 12. *If $x^{[1]}$ is strictly positive, then the iterates of the* SMART *algorithm* (203) *satisfy*

$$\text{CE}\left(x^{[k]}\right) - \text{CE}\left(x^{[k+1]}\right) \geqslant \text{KL}\left(x^{[k]}, x^{[k+1]}\right).$$

Note the difference with the first monotonicity property (182) for the Shepp–Vardi EM algorithm. The second monotonicity property is equally simple, but the precise form must be guessed. (Actually, it follows from the proof.)

Lemma 13. *If x^* is a solution of the nonnegatively constrained least-squares problem* (207)*, then, with $x^{[1]}$ strictly positive,*

$$\text{KL}\left(x^*, x^{[k]}\right) - \text{KL}\left(x^*, x^{[k+1]}\right) \geqslant \text{CE}\left(x^{[k+1]}\right) - \text{CE}\left(x^*\right) \geqslant 0.$$

Proof. Observe that

$$\begin{aligned}
&\text{KL}\left(x^*, x^{[k]}\right) - \text{KL}\left(x^*, x^{[k+1]}\right) \\
&= \sum_{p=1}^{\ell} \left\{ x_p^* \log \frac{x_p^{[k+1]}}{x_p^{[k]}} + x_p^{[k]} - x_p^{[k+1]} \right\} \\
&= \sum_{p=1}^{\ell} \left(x_p^{[k]} - x_p^* \right) \log \frac{x_p^{[k]}}{x_p^{[k+1]}} - \sum_{p=1}^{\ell} \left\{ x_p^{[k]} \log \frac{x_p^{[k]}}{x_p^{[k+1]}} + x_p^{[k+1]} - x_p^{[k]} \right\}
\end{aligned} \tag{204}$$

The last sum equals $\text{KL}(x^{[k]}, x^{[k+1]})$, and by Lemma 12, then

$$-\text{KL}(x^{[k]}, x^{[k+1]}) \geqslant \text{CE}\left(x^{[k+1]}\right) - \text{CE}\left(x^{[k]}\right).$$

The first sum equals

$$\sum_{p=1}^{\ell} \left(x_p^{[k]} - x_p^*\right) \left[A^T \log \frac{Ax^{[k]}}{b}\right]_p = \langle x^{[k]} - x^*\,,\, \nabla \text{Œ}(x^{[k]})\,\rangle,$$

where $\langle\,\cdot\,,\,\cdot\,\rangle$ denotes the inner product on $\mathbb{R}^\ell$ and ∇Œ is the gradient of Œ(x). By the convexity of Œ, then

$$\left\langle x^{[k]} - x^*\,,\, \nabla \text{Œ}\left(x^{[k]}\right)\right\rangle \geqslant \text{Œ}\left(x^{[k]}\right) - \text{Œ}\left(x^*\right).$$

Summarizing, the above shows that

$$\begin{aligned}\text{KL}\left(x^*, x^{[k]}\right) - \text{KL}\left(x^*, x^{[k+1]}\right) &\geqslant \text{Œ}\left(x^{[k]}\right) - \text{Œ}\left(x^*\right) + \text{Œ}\left(x^{[k+1]}\right) - \text{Œ}\left(x^{[k]}\right) \\ &= \text{Œ}\left(x^{[k+1]}\right) - \text{Œ}\left(x^*\right).\end{aligned}$$

This is the lemma. ■

As before, the convergence of the SMART algorithm follows starting from any strictly positive vector $x^{[1]}$.

Minimizing Burg's entropy: Compared with the minimum cross-entropy problem from the previous section, the case of Burg's entropy is problematic. For the positive system

$$Ax = b$$

with the normalization (126), the minimum Burg entropy problem is

$$\begin{aligned}&\text{minimize B E}(x) \stackrel{\text{def}}{=} \sum_{j=1}^{m} \left\{-\log \frac{b_j}{[Ax]_j} + \frac{b_j}{[Ax]_j}\right\} \\ &\text{subject to } x \geqslant 0.\end{aligned} \tag{205}$$

The first thing one notices is that it is not a convex problem. Then, it is conceivable that the solution set is not convex, and so a "second" monotonicity property is not likely to hold. However, there is a majorizing function, which suggests a multiplicative algorithm, and there is a "first" monotonicity property.

Lemma 14. *For all nonnegative x and y,*

$$BE(x) \leqslant BE(y) + \sum_{p=1}^{\ell} \left\{[A^T q]_p - [A^T r]_p \frac{y_p}{x_p}\right\} (x_p - y_p),$$

where

$$r_j = \frac{b_j}{\left([Ay]_j\right)^2}, \quad q_j = \frac{1}{[Ay]_j}, \quad j = 1, 2, \ldots, m.$$

The algorithm suggested by the lemma comes about by minimizing the upper bound on $\mathrm{BE}(x)$ with $y = x^{[k]}$, the current guess for the solution. This gives the minimizer as $x = x^{[k+1]}$,

$$x_p^{[k+1]} = x_p^{[k]} \cdot \left\{ \frac{[A^T r^{[k]}]_p}{[A^T q^{[k]}]_p} \right\}^{1/2}, \tag{206}$$

where

$$r_j^{[k]} = \frac{b_j}{\left(\left[Ax^{[k]}\right]_j\right)^2}, \quad q_j^{[k]} = \frac{1}{\left[Ax^{[k]}\right]_j}, \quad j = 1, 2, \ldots, m.$$

The "first" monotonicity property reads as follows.

Lemma 15. *Starting with a strictly positive initial guess $x^{[1]}$, the iterates generated by (206) satisfy*

$$\mathrm{BE}\left(x^{[k]}\right) - \mathrm{BE}\left(x^{[k+1]}\right) \geqslant \sum_{p=1}^{\ell} \left[A^T q^{[k]}\right]_p \frac{\left| x_p^{[k+1]} - x_p^{[k]} \right|^2}{x_p^{[k]}}.$$

It follows that the objective function decreases as the iteration proceeds, unless one has a fixed point of the iteration. It would seem reasonable to conjecture that one then gets convergence of the iterates to a local minimum, but in the absence of a second monotonicity property, this is where it ends.

Some reconstructions from simulated and real data are shown in [15]. The proofs of the above two lemmas are shown there as well.

Nonnegative Least Squares

The absence of EM algorithms for least-squares problems sooner or later had to be addressed. Here, consider positive least-squares problems, and as in section "Minimum Cross-Entropy Problems," one may as well consider them for the discrete emission tomography case. Thus, the interest is in solving the problem

$$\text{minimize} \quad \mathrm{LS}(x) \stackrel{\text{def}}{=} \| Ax - b \|^2 \quad \text{subject to} \quad x \geqslant 0. \tag{207}$$

Recall the properties (126) and (127) of the nonnegative matrix $A \in \mathbb{R}^{m \times \ell}$, and that b is a nonnegative vector. It is useful to define the operator T on nonnegative vectors by

$$[Ty]_p = y_p \frac{[A^T b]_p}{[A^T Ay]_p}, \quad p = 1, 2, \ldots, \ell. \tag{208}$$

The following discussion of the convergence of this algorithm follows De Pierro [28] and Eggermont [33]. The first item on the agenda is to prove an analogue of Lemma 1.

Lemma 16. *For all nonnegative x, $y \in \mathbb{R}^\ell$, with y strictly positive,*

$$\mathrm{LS}(x) \leqslant \mathrm{LS}(y) + \sum_{p=1}^{\ell} [A^T b]_p \left\{ \frac{(x_p - [Ty]_p)^2}{[Ty]_p} - \frac{(y_p - [Ty]_p)^2}{[Ty]_p} \right\}.$$

Proof. Observe that

$$\mathrm{LS}(x) = \mathrm{LS}(y) + 2 \langle x - y, A^T(Ay - b) \rangle + \| A(x - y) \|^2.$$

Let $z = x - y$. Write $Az = A\{y^{1/2}(z/y^{1/2})\}$ (with componentwise vector operations) and use Cauchy–Schwarz. Then,

$$\| Az \|^2 \leqslant \sum_{j=1}^{m} [Ay]_j [A(z^2/y)]_j = \sum_{p=1}^{\ell} z_p^2 \frac{[A^T Ay]_p}{y_p}.$$

Now, consider $\| Az \|^2 + 2 \langle z, A^T(Ay - b) \rangle$. Completing the square gives

$$\| Az \|^2 + 2 \langle z, A^T(Ay - b) \rangle \leqslant \sum_{p=1}^{\ell} \frac{[A^T Ay]_p}{y_p} \left(z_p + y_p \frac{[A^T(Ay - b)]_p}{[A^T Ay]_p} \right)^2 - \sum_{p=1}^{\ell} \frac{y_p}{[A^T Ay]_p} [A^T(Ay - b)]_p^2. \tag{209}$$

For the first sum, note that the expression inside the parentheses equals $x_p - [Ty]_p$. Also, $[A^T Ay]_p / y_p = [A^T b]_p / [Ty]_p$, so that takes care of the first sum. For the second sum, note that

$$\frac{y_p}{[A^T Ay]_p} [A^T(Ay - b)]_p^2 = \frac{[A^T Ay]_p}{y_p} \left(y_p - y_p \frac{[A^T b]}{[A^T Ay]_p} \right)^2,$$

and the expression for the second sum follows. ∎

It is now clear how one may construct an algorithm. Take $y = x^{[1]}$, a strictly positive vector, and minimize the upper bound on $\mathrm{LS}(x)$ given by the lemma. Setting the gradient equal to 0 gives $x^{[k+1]} = T x^{[k]}$ or

$$x_p^{[k+1]} = x_p^{[k]} \cdot \frac{[A^T b]_p}{[A^T A x^{[k]}]_p}, \quad p = 1, 2, \ldots, \ell. \tag{210}$$

Note that then all $x^{[k]}$ are strictly positive because A is nonnegative and has unit column sums.

This algorithm is due to Daube-Witherspoon and Muehllehner [26] for emission tomography with the acronym ISRA.

Onward to the monotonicity properties of the algorithm, the following lemma is immediate.

Lemma 17. *If $x^{[1]}$ is strictly positive, then the iterates of the* ISRA *algorithm* (210) *satisfy*

$$\mathrm{LS}\left(x^{[k]}\right) - \mathrm{LS}\left(x^{[k+1]}\right) \geqslant \sum_{p=1}^{\ell} [A^T b]_p \, \frac{\left(x_p^{[k]} - x_p^{[k+1]}\right)^2}{x_p^{[k+1]}}.$$

The "second" monotonicity property is a bit more involved than for the EM algorithm but still involves Kullback–Leibler distances. Let

$$\mathrm{KLS}(x, y) = \mathrm{KL}(c \cdot x, c \cdot y) + \mathrm{LS}(y) - \mathrm{LS}(y^*), \tag{211}$$

where $x = y^*$ is any minimizer of $\| Ax - b \|^2$ over $x \geqslant 0$. Here, $c = A^T b$ and the dot means componentwise multiplication.

Lemma 18. *If x^* is a solution of the nonnegatively constrained least-squares problem* (207)*, then*

$$\mathit{KLS}(c \cdot x^*, c \cdot x^{[k]}) - \mathit{KLS}(c \cdot x^*, c \cdot x^{[k+1]}) \geqslant \frac{1}{2} \mathrm{LS}(x^{[k]}) - \frac{1}{2} \mathrm{LS}(x^*) \geqslant 0.$$

Proof. As before, one has

$$\begin{aligned}
&\mathrm{KL}(c \cdot x^*, c \cdot x^{[k]}) - \mathrm{KL}(c \cdot x^*, c \cdot x^{[k+1]}) \\
&\qquad = \sum_{p=1}^{\ell} c_p \, x_p^* \log \frac{x_p^{[k+1]}}{x_p^{[k]}} + c_p \left(x_p^{[k]} - x_p^{[k+1]}\right) \\
&\qquad \geqslant \sum_{p=1}^{m} c_p \, x_p^* \left(1 - \frac{x_p^{[k]}}{x_p^{[k+1]}}\right) + c_p \left(x_p^{[k]} - x_p^{[k+1]}\right) \\
&\qquad \geqslant \sum_{p=1}^{\ell} \frac{c_p \left(x_p^{[k+1]} - x_p^*\right)\left(x_p^{[k]} - x_p^{[k+1]}\right)}{x_p^{[k+1]}}.
\end{aligned} \tag{212}$$

Here, in the second line, the inequality $\log t = -\log(t^{-1}) \geqslant 1 - t^{-1}$ was used.

Now, let $C \in \mathbb{R}^{\ell \times \ell}$ be the diagonal matrix with diagonal components

$$C_{p,p} = \frac{[A^T b]_p}{x_p^{[k+1]}}, \quad p = 1, 2, \ldots, \ell.$$

Then, the least expression equals

$$\langle\, x^{[k+1]} - x^*\,,\, C\,(x^{[k]} - x^{[k+1]})\,\rangle =$$
$$\langle\, x^{[k+1]} - x^{[k]}\,,\, C\,(x^{[k]} - x^{[k+1]})\,\rangle + \langle\, x^{[k]} - x^*\,,\, C\,(x^{[k]} - x^{[k+1]})\,\rangle.$$

Since $C\,(x^{[k]} - x^{[k+1]}) = A^T(Ax^{[k]} - b)$, then

$$\langle\, x^{[k]} - x^*\,,\, C\,(x^{[k]} - x^{[k+1]})\,\rangle = \langle\, x^{[k]} - x^*\,,\, A^T(Ax^{[k]} - b)\,\rangle$$
$$\geqslant \frac{1}{2}\,\mathrm{LS}(x^{[k]}) - \frac{1}{2}\,\mathrm{LS}(x^*),$$

the last inequality by convexity. Finally,

$$\langle\, x^{[k+1]} - x^{[k]}\,,\, C\,(x^{[k]} - x^{[k+1]})\,\rangle = -\sum_{p=1}^{\ell} \frac{[A^T b]_p}{x_p^{[k+1]}} \left(x_p^{[k]} - x_p^{[k+1]}\right)^2.$$

which by Lemma 17 dominates $\mathrm{LS}(x^{[k+1]}) - \mathrm{LS}(x^{[k]})$. This shows that

$$\mathrm{KL}(c \cdot x^*, c \cdot x^{[k]}) - \mathrm{KL}(c \cdot x^*, c \cdot x^{[k+1]}) \geqslant \frac{1}{2}\,\mathrm{LS}(x^{[k]}) - \frac{1}{2}\,\mathrm{LS}(x^*)$$
$$- \mathrm{LS}(x^{[k]}) + \mathrm{LS}(x^{[k+1]}).$$

and the lemma follows. ■

The convergence of the algorithm now follows similar to the EM case.

Multiplicative Iterative Algorithms

This final section concerns the observation that multiplicative iterative algorithms may be constructed by way of proximal point algorithms as in [33] and that Kullback–Leibler distances naturally appear in this context. For arbitrary convex functions F on $\mathbb{R}^\ell$, one may solve the problem with nonnegativity constraints

$$\text{minimize} \quad F(x) \quad \text{subject to} \quad x \geqslant 0 \tag{213}$$

by computing a sequence $\{x^{[k]}\}_k$, with $x^{[k+1]}$ the solution of

$$\text{minimize} \quad F(x) + (\omega_k)^{-1}\,\mathrm{KL}(x^{[k]}, x) \quad \text{subject to} \quad x \geqslant 0, \tag{214}$$

starting from some $x^{[1]}$ with strictly positive components. Here, $\omega_k > 0$. One verifies that $x^{[k+1]}$ satisfies

$$x_p^{[k+1]} = \frac{1}{1 + \omega_k\,[\nabla F(x^{[k+1]})]_p}, \quad p = 1, 2, \ldots, \ell. \tag{215}$$

This is an implicit equation for $x^{[k+1]}$, but explicit versions suggest themselves. Note that the objective function in (214) is strictly convex, so that the solution is unique, assuming solutions exist. Of course, other proximal functions suggest themselves, such as $\mathrm{KL}(x, x^{[k]})$. See, e.g., [19]. The classical one is $\| x - x^{[k]} \|^2$, the squared Euclidean distance, due to Rockafellar. See, e.g., [13].

It is interesting to note that the implicit algorithm (215) satisfies the two monotonicity properties. The first one is obvious,

$$F(x^{[k]}) - F(x^{[k+1]}) \geqslant (\omega_k)^{-1}\,\mathrm{KL}(x^{[k]}, x^{[k+1]}), \tag{216}$$

since $x^{[k+1]}$ is the minimizer of (214).

For the second monotonicity property, assume that x^* is a solution of (213). Note that

$$\begin{aligned}
\mathrm{KL}(x^*, x^{[k]}) - \mathrm{KL}(x^*, x^{[k+1]}) &= \sum_{p=1}^{\ell} x_p^* \log \frac{x_p^{[k+1]}}{x_p^{[k]}} + x_p^{[k]} - x_p^{[k+1]} \\
&= \sum_{p=1}^{\ell} x^* \log \frac{1}{1 + \omega_k\,[\nabla F(x^{[k+1]})]_p} \\
&\quad + \omega_k\, x^{[k+1]}\,[\nabla F(x^{[k+1]})]_p \\
&\geqslant \sum_{p=1}^{\ell} -x_p^*\,\omega_k\,[\nabla F(x^{[k+1]})]_p + \omega_k\, x^{[k+1]}\,[\nabla F(x^{[k+1]})]_p \\
&= \omega_k \langle\, x^{[k+1]} - x^*, \nabla F(x^{[k+1]})\rangle \geqslant \omega_k\, \left(F(x^{[k+1]}) - F(x^*)\right).
\end{aligned}$$

To summarize,

$$\mathrm{KL}(x^*, x^{[k]}) - \mathrm{KL}(x^*, x^{[k+1]}) \geqslant \omega_k\, \left(F(x^{[k+1]}) - F(x^*)\right). \tag{217}$$

The convergence of the algorithm follows. Practically speaking, one has to devise explicit versions of the algorithm and see how they behave.

10 Accelerating the EM Algorithm

The Ordered Subset EM Algorithm

It is well known that EM algorithms converge very slowly, even if one wants to stop the iteration "early." For the general EM algorithm of section "The Bare-Bones EM Algorithm Fleshed Out," some attempts at acceleration have been made along the lines of coordinate descent methods or more generally descent along groups of coordinates. In particular, the M-step (34) is replaced by a sequence of M-steps, for $j = 1, 2, \ldots, m$, with $x_1 = x^{[k]}$,

$$\begin{aligned} &\text{minimize} \quad \mathfrak{L}(x|x_{j-1}) \stackrel{\text{def}}{=} -\int_{\mathcal{Z}} \varphi_{\mathcal{Z}}(z|x_{j-1}) \log f_{\mathcal{Z}}(z|x)\, d\mu(z) \\ &\text{subject to} \quad x \in \mathcal{X}_j, \end{aligned} \tag{218}$$

and then $x^{[k+1]} = x_m$. Here $\{\mathcal{X}_j\}_{j=1}^m$ is a not necessarily disjoint division of the parameter space $\mathcal{X}$. See [68] and references therein. In the context of emission tomography, the more generally accepted route to accelerating the EM algorithm has been via the ordered subset approach of Hudson and Larkin [59]. Without putting too fine a point to it, this amounts to partitioning the data space rather than the parameter space. The acceleration achieved by these methods seems to be twofold. The ordered subset approach allows for more efficient computer implementations *and*, the convergence itself is speeded up. See, e.g., [60].

The ordered subset EM algorithm (OSEM) of Hudson and Larkin [59] deals with the maximum likelihood problem of emission tomography (130). The starting point is to divide the data into blocks, characterized by the sets of indices $\Omega(1), \Omega(2), \ldots, \Omega(s)$ such that

$$\Omega(1) \cup \Omega(2) \cup \ldots \cup \Omega(s) = \{1, 2, \ldots, m\}. \tag{219}$$

However, the sets need not be disjoint. Define the partial negative Kullback–Leibler functionals

$$L_r(x) = \sum_{j \in \Omega(r)} \left\{ b_j \log \frac{b_j}{[Ax]_j} + [Ax]_j - b_j \right\}, \quad r = 1, 2, \ldots, s. \tag{220}$$

Note that for all r,

$$\sum_{j \in \Omega(r)} [Ax]_j = \sum_{p=1}^{\ell} \alpha_{rp}\, x_p \quad \text{with} \quad \alpha_{rp} = \sum_{j \in \Omega(r)} a(j, p). \tag{221}$$

The OSEM algorithm now consists of successively applying one step of the Shepp–Vardi EM algorithm to each of the problems

$$\text{minimize} \quad L_r(x) \quad \text{subject to} \quad x \geqslant 0. \tag{222}$$

To spell out the OSEM iteration exactly, it is useful to introduce the data vectors B_r and the matrices A_r by

$$B_r = \big(b_j : j \in \Omega(r)\big) \quad \text{and} \quad A_r x = \big([Ax]_j : j \in \Omega(r)\big). \tag{223}$$

Then, $L_r(x) = \mathrm{KL}(B_r, A_r x)$, and the OSEM algorithm takes the form

$$x_p^{[k+1]} = x_p^{[k]} \cdot \alpha_{rp}^{-1} \, [A_r^T \varrho^{[k]}]_p, \quad p = 1, 2, \ldots, \ell, \tag{224}$$

where $r = k \bmod s$ (in the range $1 \leqslant r \leqslant s$) and $\varrho_q^{[k]} = B_{rq} \big/ [A_r x^{[k]}]_q$ for all q. The slight complication of the α_{rp} arises because the matrices A_r do not have unit column sums. This is fixed by defining the matrices $\mathbb{A}_r$ by

$$[\mathbb{A}_r]_{qp} = \alpha_{rp}^{-1} \, [A_r]_{qp} \quad \text{for all} \quad q \text{ and } p. \tag{225}$$

Now, define $\mathbb{L}_r(y) = L_r(x) = \mathrm{KL}(B_r, \mathbb{A}_r y)$, where $y_q = \alpha_{rq} x_q$ for all q. A convenient short-hand notation for this is $y = \alpha_r \cdot x$. Now, if x minimizes $L_r(x)$, then $y = \alpha_r \cdot x$ minimizes $\mathbb{L}_r(y)$ and vice versa. Since the matrices $\mathbb{A}_r$ have unit column sums, the EM algorithm for minimizing $\mathbb{L}_r(y)$ is

$$y_p^{[k+1]} = y_p^{[k]} \cdot [\mathbb{A}_r^T \varrho^{[k]}]_p \,, \quad p = 1, 2, \ldots, \ell, \tag{226}$$

with $\varrho_q^{[k]} = B_{rq} \big/ [\mathbb{A}_r y^{[k]}]_q$ for all q. Transforming back gives (224).

Regarding the convergence of the OSEM algorithm, the best one can hope for is cyclic convergence, i.e., each of the subsequences $\big\{x^{[k+r\,s]}\big\}_{r \geqslant 1}$ converges. Proving this would be a daunting task. However, as observed by Byrne [9], it is useful to consider what happens if the system of equations $Ax = b$ is consistent in the sense that

$$\exists \, x^* \geqslant 0 \; : \; Ax^* = b, \tag{227}$$

when one should expect convergence of the whole sequence to a nonnegative solution of $Ax = b$. Hudson and Larkin [59] prove that this is so under the so-called subset-balancing condition

$$\alpha_{rp} = \alpha_{o,p}, \quad p = 1, 2, \ldots, \ell \quad \text{and} \quad r = 1, 2, \ldots, s. \tag{228}$$

That is, the column sums are the same for all blocks. This is a strong condition, even if one allows for overlapping blocks of data. Haltmeier et al. [51] make the same assumption in the continuous setting. Byrne [11] observed that the condition may be relaxed to that of *subset-separability*: There exist coefficients β_r and γ_p such that

$$\alpha_{rp} = \beta_r \gamma_p, \quad r = 1, 2, \ldots, s \quad \text{and} \quad p = 1, 2, \ldots, \ell. \tag{229}$$

The convergence proof of the OSEM algorithm (224) under the subset-separability condition (229) relies on the two monotonicity properties of the EM algorithm (226); see section "Monotonicity of the Shepp–Vardi EM Algorithm." After translation, one gets the following monotonicity properties for (224).

Lemma 19. *Let x^* be a nonnegative solution of $Ax = b$. Starting from a strictly positive $x^{[1]} \in V_\ell$, then, with $r = k \bmod s$,*

$$L_r(x^{[k]}) - L_r(x^{[k+1]}) \geqslant \mathrm{KL}(\alpha_r \cdot x^{[k+1]}, \alpha_r \cdot x^{[k]}) \geqslant 0 \quad \textit{and}$$
$$\mathrm{KL}(\alpha_r \cdot x^*, \alpha_r \cdot x^{[k]}) - \mathrm{KL}(\alpha_r \cdot x^*, \alpha_r \cdot x^{[k+1]}) \geqslant L_r(x^{[k]}) - L_r(x^*) \geqslant 0.$$

Now, if the α_{rp} change with r, then one cannot conclude much from the lemma. However, under the subset-separability condition (229), one obtains for all nonnegative x and y

$$\mathrm{KL}(\alpha_r \cdot x, \alpha_r \cdot y) = \beta_r \, \mathrm{KL}(\gamma \cdot x, \gamma \cdot y),$$

and the inequalities of the lemma translate as follows.

Corollary 1. *Under the conditions of Lemma 19 and the subset-separability condition (229),*

$$\beta_r^{-1} \left\{ L_r(x^{[k]}) - L_r(x^{[k+1]}) \right\} \geqslant \mathrm{KL}(\gamma \cdot x^{[k+1]}, \gamma \cdot x^{[k]}) \geqslant 0 \quad \textit{and}$$
$$\mathrm{KL}(\gamma \cdot x^*, \gamma \cdot x^{[k]}) - \mathrm{KL}(\gamma \cdot x^*, \gamma \cdot x^{[k+1]}) \geqslant \beta_r^{-1} \left\{ L_r(x^{[k]}) - L_r(x^*) \right\} \geqslant 0.$$

As in Theorem 1, one may conclude that the sequence $\{\gamma \cdot x^{[k]}\}_k$ converges to a nonnegative solution x^{**} of $Ax = b$. Note that there is another way of looking at this; see Remark 5 at the end of this chapter.

As remarked, the subset-separability condition (228) is very strong. It fails dramatically in the extreme case of the OSEM algorithm when each block consists of a single row. In that case, the OSEM algorithm (224) reduces to

$$x_p^{[k+1]} = x_p^{[k]} \cdot \frac{b_j}{[Ax^{[k]}]_j}, \quad p = 1, 2, \ldots, \ell, \tag{230}$$

where $j = k \bmod m$. So, $x^{[k+1]}$ is a multiple of $x^{[k]}$, and the OSEM algorithm produces only multiples of the initial guess $x^{[1]}$. So, certainly in this case, the algorithm does not converge, but more seriously, it does not do anything useful.

So, what is one to do? Following Byrne [9] (see also [13]), the next section turns to row-action methods (where the blocks consist of a single datum), that is, the (additive) algebraic reconstruction technique (ART) and the multiplicative version (MART) of Gordon et al. [48], as well as block-iterative variants. (The

simultaneous version (SMART) of the multiplicative version was already discussed in section "Minimum Cross-Entropy Problems.") This points into the direction of relaxation and scaling. After that, the table is set for block-iterative versions of the EM algorithm.

The ART and Cimmino–Landweber Methods

It is useful to discuss the situation in regard to the so-called *algebraic reconstruction technique* (ART) of Gordon et al. [48], and the Cimmino–Landweber iteration, a reasonable version of the original SIRT method. Herman [55] is the authoritative source, but see [95] for a comparison with conjugate gradient method. The ART and Cimmino–Landweber algorithms were designed to solve systems of linear equations of the form

$$A\,x = b \tag{231}$$

with b the measured nonnegative data and $A \in \mathbb{R}^{m\times\ell}$ with nonnegative components $a(j,\,p)$ but not necessarily unit column sums. Of course, for inconsistent systems of equations, this must be replaced by the least-squares problem

$$\text{minimize} \quad \| Ax - b \|^2 \quad \text{subject to} \quad x \in \mathbb{R}^m, \tag{232}$$

but in fact, the ART method solves the weighted least-squares problem

$$\text{minimize} \quad \sum_{j=1}^{m} \frac{\left|\left\langle a(j,\,\cdot), x\right\rangle - b_j\right|^2}{\| a(j,\,\cdot)\|^2} \quad \text{subject to} \quad x \in \mathbb{R}^m, \tag{233}$$

A standard method for the solution of (232) is the Cimmino–Landweber iteration

$$x^{[k+1]} = x^{[k]} + \omega_k\, A^T(b - Ax^{[k]}), \tag{234}$$

for suitably small but not too small positive relaxation parameters ω_k. It is mentioned here for its analogy with the EM algorithm. The Cimmino–Landweber iteration is a well-studied method for the regularization of the least-squares problem (232) and is itself subject to acceleration; see, e.g., Hanke [52] and references therein.

At the other end of the spectrum is Kaczmarz' method, which consists of sequential orthogonal projections onto the hyperplanes

$$H_j = \left\{x \in \mathbb{R}^\ell \,\middle|\, \left\langle a(j,\,\cdot\,),\, x\right\rangle = b_j\right\}.$$

Formally, this is achieved by computing the new iterate $x^{[k+1]}$ from the previous one $x^{[k]}$ by solving

$$\text{minimize} \quad \| x - x^{[k]} \|^2 \quad \text{subject to} \quad \langle a(j, \cdot), x \rangle = b_j. \tag{235}$$

The iteration then takes the form, with $j = k \bmod m$,

$$x^{[k+1]} = x^{[k]} + \omega_k \frac{b_j - \langle a(j, \cdot), x^{[k]} \rangle}{\| a(j, \cdot) \|^2} a(j, \cdot) \tag{236}$$

for $\omega_k = 1$. The relaxation parameter ω_k is included to see whether choices other than $\omega_k = 1$ might be advantageous. Geometrically, a requirement is $0 < \omega_k < 2$. The choice $\omega_k = 0$ would not do anything; the choice $\omega_k = 2$ implements reflection with respect to the hyperplane H_j. The algorithm (236) with relaxation originated with Gordon et al. [48].

Typically, one takes the hyperplanes in cyclic order $j = k \bmod m$, but Herman and Meyer [56] show experimentally that carefully reordering the hyperplanes has a big effect on the quality of the reconstruction when the number of iterations is fixed before hand. The choice of $\omega (= \omega_k$ for all $k)$ also matters greatly, but the optimal one seems to depend on everything (the experimental setup leading to the matrix A, the noise level, etc.), so that the optimal ω can be very close to 0 or close to 2 or in between.

Byrne [8, 9] observes that the scaling of the ART algorithm is just about optimal, as follows. Actually, it is difficult to say much for inconsistent systems, other than experimentally (see [56]), but for consistent systems, one has the following two monotonicity properties, which are reminiscent of the monotonicity properties for the Shepp–Vardi EM algorithm. However, they are much less impressive since they only hold for consistent systems. Define

$$\mathrm{LS}_j(x) = \frac{|\langle a(j, \cdot), x \rangle - b_j|^2}{\| a(j, \cdot) \|^2}, \quad j = 1, 2, \ldots, m. \tag{237}$$

Lemma 20. *If x^* satisfies $Ax^* = b$, then*

$$\mathrm{LS}_j(x^{[k]}) - \mathrm{LS}_j(x^{[k+1]}) = \omega_k(2 - \omega_k)\,\mathrm{LS}_j(x^{[k]}),$$
$$\| x^{[k]} - x^* \|^2 - \| x^{[k+1]} - x^* \|^2 = \omega_k(2 - \omega_k)\,\mathrm{LS}_j(x^{[k]}).$$

The proofs involve only (exact) quadratic Taylor expansions and are omitted. The conclusion is that ART converges in the consistent case if $\omega_k = \omega$ is constant and $0 < \omega < 2$. Following Byrne [8], one notes that the second monotonicity property suggests that $\omega_k(2 - \omega_k)$ should be as large as possible. This is achieved by $\omega_k = 1$. In other words, the original Kaczmarz procedure (236) with $\omega_k = 1$ is optimally scaled. However, as already remarked above, a choice other than $\omega_k = 1$ may speed things up initially.

Despite the good news that ART is much faster than the Cimmino–Landweber type methods, it is still "slow." Now, in transmission tomography as in emission

tomography, the system of equations $Ax = b$ naturally decomposes into a number of blocks

$$A_r x = B_r, \quad r = 1, 2, \ldots, s, \tag{238}$$

(see (223)), and then one has the block version of (235)

$$\text{minimize} \quad \| x - x^{[k]} \|^2 \quad \text{subject to} \quad A_r x = B_r, \tag{239}$$

with the solution

$$x^{[k+1]} = x^{[k]} + \omega_k \, A_r^T \left(A_r A_r^T \right)^\dagger (B_r - A_r x^{[k]}), \tag{240}$$

where $\dagger$ denotes the Moore–Penrose inverse. Now, computing $\left(A_r^T A_r \right)^\dagger w$ (for any vector w) would be expensive, but it seems reasonable that $A_r^T A_r$ should be close to diagonal, in which case one may just replace it with its diagonal. This leads to the algorithm

$$x^{[k+1]} = x^{[k]} + \omega_k \, A_r^T D_r^{-1} (B_r - A_r x^{[k]}), \tag{241}$$

where D_r is a diagonal matrix with $[D_r]_{qq} = \left[A_r A_r^T \right]_{qq}$.

Now, it turns out that computing $A_r x$ is not much more expensive than computing a single $\langle a(j, \cdot), x \rangle$ and that the matrices $A_r^T A_r$ are very close to diagonal, so that one step of the block method (241) practically achieves as much as the combined ART steps for all the equations in one block. So, methods that process naturally ordered blocks are appreciably faster than the two extreme methods. See [35].

It is not clear how to choose the optimal relaxation parameters. Regarding (241), it is known that the algorithm converges cyclically provided the blocks and the relaxation parameters are chosen cyclically, i.e., if $r = k \bmod s$ and $\omega_k \equiv \omega_r$, and

$$\max_{1 \leqslant r \leqslant s} \left\| I - \omega_r A_r D_r^{-1} A_r^T \right\|_2 < 1, \tag{242}$$

then $\{x^{[r+ks]}\}_k$ converges for each $r = 1, 2, \ldots, s$; see [35]. Moreover, if the relaxation parameter is kept fixed, say,

$$\omega_k = \omega \quad \text{for all} \quad k, \tag{243}$$

and denoting the iterates by $x^{[i+kI]}(\omega)$ to show the dependence on ω, then

$$\lim_{\omega \to 0} \lim_{k \to \infty} x^{[i+kI]}(\omega) = x^*, \tag{244}$$

the minimum norm solution of (233), provided the initial guess belongs to the range of A^T. See [16]. At about the same time, Trummer [94] showed for the relaxed ART method (236) that

$$\lim_{k\to\infty} x^{[k]}(\omega_k) = x^*, \tag{245}$$

provided

$$\omega_k > 0, \quad \sum_{k=1}^{\infty} \omega_k^2 < \infty \quad \text{and} \quad \sum_{k=1}^{\infty} \omega_k = +\infty. \tag{246}$$

Note the difference between (245) and (246).

The MART and SMART Methods

Consider again the system of equations $Ax = b$ as it arises in the PET setting, with A and b having nonnegative components and A having unit column sums. In section "Minimum Cross-Entropy Problems," the SMART algorithm was discussed for the solution of

$$\text{minimize} \quad \mathrm{KL}(Ax, b) \quad \text{subject to} \quad x \geqslant 0, \tag{247}$$

i.e.,

$$x_p^{[k+1]} = x_p^{[k+1]} \cdot \exp\left(\left[A^T \log \frac{b}{Ax^{[k]}}\right]_p\right), \quad p = 1, 2, \ldots, \ell. \tag{248}$$

The multiplicative ART algorithm (MART) of Gordon et al. [48] formally arises as the multiplicative version of the additive ART algorithm, to wit

$$x_p^{[k+1]} = x_p^{[k]} \cdot \left(\frac{b_j}{\left\langle a(j,\cdot), x^{[k]}\right\rangle}\right)^{a(j,p)}, \quad p = 1, 2, \ldots, \ell$$

or equivalently for $p = 1, 2, \ldots, \ell$

$$x_p^{[k+1]} = x_p^{[k]} \cdot \exp\left(\omega_k\, a(j,p) \log \frac{b_j}{\left\langle a(j,\cdot), x^{[k]}\right\rangle}\right) \tag{249}$$

with $\omega_k = 1$. Again, the relaxation parameter ω_k was included to explore whether choices other than $\omega_k = 1$ would be advantageous. Byrne [9] observes that the MART algorithm typically does not enjoy the same speedup compared to the simultaneous SMART version that ART has over Cimmino–Landweber. To get some insight into this, it is useful to consider a projection method analogous to the Kaczmarz method of orthogonal projections onto hyperplanes. The method in question is well known (see, e.g., [17]),

$$\text{minimize} \quad \mathrm{KL}(x, x^{[k]}) \quad \text{subject to} \quad \left\langle a(j,\cdot), x^{[k]}\right\rangle = b_j. \tag{250}$$

One may approximately solve this as follows. With the unrestricted Lagrange multiplier λ, one gets the equations $\log\left(x_p/x_p^{[k]}\right) + \lambda a(j,p) = 0$ for all p, so that

$$x_p = x_p^{[k]} \exp\left(\lambda\, a(j,p)\right), \quad p = 1, 2, \ldots, \ell.$$

To enforce the constraint, take inner products with $a(j,\cdot)$. This results in

$$b_j = \left\langle a(j,\cdot), x \right\rangle = \sum_{p=1}^{\ell} a(j,p)\, x_p^{[k]} \exp\left(\lambda\, a(j,p)\right), \tag{251}$$

and one would like to solve this for λ. That does not appear manageable, but it can be done approximately as follows. Since the $a(j,p)$ and $x_p^{[k]}$ are nonnegative, by the mean value theorem, there exists a θ, with

$$0 < \theta < \max\left\{a(j,p) : 1 \leqslant p \leqslant \ell\right\}, \tag{252}$$

such that the right-hand side of (251) equals $\exp(\lambda\,\theta)\left\langle a(j,\cdot), x^{[k]}\right\rangle$. Then, solving (251) for λ gives the iteration

$$x_p^{[k+1]} = x_p^{[k]} \exp\left(\omega\, a(j,p)\, \log \frac{b_j}{\left\langle a(j,\cdot), x^{[k]}\right\rangle}\right), \tag{253}$$

with $\omega = 1/\theta$. The conservative choice, the one that changes $x^{[k]}$ the least, is to choose ω as small as possible. In view of (252), this gives $\omega = \omega_j$,

$$\omega_j = \frac{1}{\max\limits_{1\leqslant p\leqslant \ell} a(j,p)}. \tag{254}$$

Note that if A has unit column sums, then one may expect ω to be quite large. This may explain why the original MART algorithm is not greatly faster than the SMART version. In defense of Gordon et al. [48], one should mention that they considered matrices with components 0 or 1, in which case $\omega = 1$!

Following Byrne [8], the block-iterative version of (253) is as follows. In the partitioned data setup of (223), the BI-MART algorithm is

$$x_p^{[k+1]} = x_p^{[k]} \cdot \exp\left(\frac{\omega_k}{\alpha_r}\left[A_r^T \log\left\{\frac{B_r}{A_r x^{[k]}}\right\}\right]_p\right) \tag{255}$$

for $p = 1, 2, \ldots, \ell$, where

$$\alpha_r = \max\left\{\alpha_{rp} : 1 \leqslant p \leqslant \ell\right\} \tag{256}$$

is the maximal column sum of A_r. One would expect that $\omega_k = 1$ should be the optimal choice. This is the rescaled BI-MART (or RBI-MART) algorithm of Byrne [8]. Following the template of section "Minimum Cross-Entropy Problems," one proves the following majorizing inequality and the two monotonicity properties. For nonnegative vectors y and $\omega > 0$, define $R_\omega y$ by

$$[R_\omega y]_p = y_p \exp\left(\frac{\omega}{\alpha_r}\left[A_r^T \log\left\{\frac{B_r}{A_r y}\right\}\right]_p\right) \tag{257}$$

for $p = 1, 2, \ldots, \ell$.

Lemma 21. *For all nonnegative x and y,*

$$\mathrm{KL}(A_r x, B_r) \leqslant \mathrm{KL}(A_r y, B_r) + \frac{\alpha_r}{\omega}\{\mathrm{KL}(x, R_\omega y) - \mathrm{KL}(y, R_\omega y)\}.$$

Note that the minimizer of the right-hand side is $x = R_\omega y$. This would give rise to the algorithm (255).

Proof. Recall the identity from section "Minimum Cross-Entropy Problems,"

$$\mathrm{KL}(A_r x, B_r) = \mathrm{KL}(A_r y, B_r) + \mathrm{KL}(A_r x, A_r y) + \left\langle x - y, A_r^T \varrho \right\rangle,$$

with $\varrho_j = \log\left([A_r y]_j \big/ [B_r]_j\right)$. Now, a convexity argument gives that

$$\mathrm{KL}(A_r x, A_r y) \leqslant \alpha_r \,\mathrm{KL}(x, y),$$

so that one gets the inequality

$$\mathrm{KL}(A_r x, B_r) \leqslant \mathrm{KL}(A_r y, B_r) + \alpha_r \,\mathrm{KL}(x, y) + \left\langle x - y, A_r^T \varrho \right\rangle.$$

The definition of the operator R_ω gives that

$$\log \frac{[R_\omega y]_p}{y_p} = -\frac{\omega}{\alpha_r} A_r^T \varrho,$$

so then, with $\theta \equiv 1/\omega$,

$$\begin{aligned}
&\alpha_r \,\mathrm{KL}(x, y) + \left\langle x - y, A_r^T \varrho \right\rangle \\
&\quad = \alpha_r \left\{\mathrm{KL}(x, y) - \theta \left\langle x - y, \log \frac{R_\omega y}{y} \right\rangle\right\} \\
&\quad = \alpha_r \,\{(1 - \theta)\,\mathrm{KL}(x, y) + \theta\,(\mathrm{KL}(x, R_\omega y) - \mathrm{KL}(y, R_\omega y))\}.
\end{aligned}$$

The last line follows after some lengthy bookkeeping. So, for $\theta \leqslant 1$ or $\omega \geqslant 1$, the conclusion follows. ∎

The first monotonicity property then follows.

Lemma 22. *For $\omega \geqslant 1$ and $r = k \bmod s$,*

$$\mathrm{KL}\left(A_r x^{[k]}, B_r\right) - \mathrm{KL}\left(A_r x^{[k+1]}, B_r\right) \geqslant \frac{\alpha_r}{\omega}\, \mathrm{KL}\left(x^{[k]}, x^{[k+1]}\right) \geqslant 0.$$

The second monotonicity property follows after some work (omitted).

Lemma 23. *If $x^* \geqslant 0$ satisfies $Ax^* = b$, then for all k and $r = k \bmod s$,*

$$\mathrm{KL}\left(x^*, x^{[k]}\right) - \mathrm{KL}\left(x^*, x^{[k+1]}\right) \geqslant \frac{\omega}{\alpha_r}\, \mathrm{KL}\left(A_r x^{[k+1]}, B_r\right) \geqslant 0. \tag{258}$$

Now, regardless of whether $\omega = 1$ maximizes the right-hand side of the inequality (258), the presence of the factor α_r, which should be small if the original matrix A has unit column sums, suggests that the choice $\omega = 1$ in (258) is a tremendous improvement over the case $\omega = \alpha_r$, which would arise if one ignored the non-unit column sums of A_r.

Row-Action and Block-Iterative EM Algorithms

Attention now turns to the construction of the row-action version of the EM algorithm and the associated block-iterative versions. Recall the formulation (130) of the maximum likelihood problem for the PET problem as

$$\text{minimize} \quad \mathrm{KL}(b, Ax) \quad \text{subject to} \quad x \in V_\ell, \tag{259}$$

with $b \in \mathbb{R}^m$ and $A \in \mathbb{R}^{m \times \ell}$ nonnegative, with A having unit column sums.

Now, construct a row-action version by considering the following iterative projection method, where the new iterate $x^{[k+1]}$ is obtained by projecting the previous iterate $x^{[k]}$ onto the hyperplane $\left\langle a(j, \cdot)\,, x \right\rangle = b_j$. The particular projection is obtained by

$$\text{minimize} \quad \mathrm{KL}(x^{[k]}, x) \quad \text{subject to} \quad \left\langle a(j, \cdot), x \right\rangle = b_j. \tag{260}$$

Again, with λ, an unrestricted Lagrange multiplier, one must solve the equations $-x_p^{[k]}/x_p + 1 + \lambda\, a(j, p) = 0$, or

$$x_p = x_p^{[k]} - \lambda\, a(j, p)\, x_p, \quad p = 1, 2, \ldots, \ell. \tag{261}$$

At this point, one must make the simplification where x_p on the right-hand side is replaced by $x_p^{[k]}$. This gives the equation

$$x_p = x_p^{[k]} - \lambda\, a(j,p)\, x_p^{[k]}, \quad p = 1, 2, \ldots, \ell.$$

To enforce the constraint, multiply by $a(j,p)$ and sum over p. Then,

$$b_j = \left\langle a(j,\cdot), x^{[k]} \right\rangle - \lambda \sum_{p=1}^{\ell} a(j,p)^2\, x_p^{[k]} = (1 - \lambda\,\theta) \left\langle a(j,\cdot), x^{[k]} \right\rangle. \tag{262}$$

where in the last line the mean value theorem was used, for some θ satisfying

$$0 < \theta < \max\left\{a(j,p) : 1 \leqslant p \leqslant \ell\right\}. \tag{263}$$

Solving for λ gives the iterative step

$$x^{[k+1]} = (1 - \omega\, a(j,p))\; x_p^{[k]} + \omega\, x_p^{[k]}\, \frac{a(j,p)\, b_j}{\left\langle a(j,\cdot), x^{[k]} \right\rangle}, \tag{264}$$

for $p = 1, 2, \ldots, \ell$, where $\omega \equiv 1/\theta$. In the notation of (223), with some imagination, the block-iterative version is then

$$x^{[k+1]} = R_{\omega,r}\, x^{[k]}, \tag{265}$$

where the operators $R_{\omega,r}$ are defined by

$$[R_{\omega,r}\, x]_p = \left(1 - \omega\, \alpha_{rp}\right)\, x_p + \omega\, x_p \left[A_r^T \left(B_r / A_r x\right)\right]_p, \tag{266}$$

for $p = 1, 2, \ldots, \ell$. So now, ω is considered to be a relaxation parameter.

The algorithm (266) was obtained by Byrne [8, 9] after carefully examining the analogy with MART vs. RBI-SMART. His choice for the relaxation parameter ω is to take it depending on the block, so $\omega = \omega_r$ with

$$\omega_r = \frac{1}{\max\limits_{1 \leqslant p \leqslant \ell} \alpha_{rp}}, \tag{267}$$

which he obtained by deriving the two monotonicity properties discussed below. Byrne [8, 9] designated the resulting algorithm (266)–(267) as rescaled block-iterative EM for maximum likelihood algorithm (RBI-EMML). At about the same time, Browne and De Pierro [5] discovered the algorithm (264)–(266). They named (264) the RAMLA (row-action maximum likelihood algorithm). For the latest on this, see [92].

The above considerations strongly suggest that algorithm (266)–(267) is the correct one. This is corroborated by practical experience. The following monotonicity properties lend even more weight to it. A slight drawback is that they require that the system $Ax = b$ has a nonnegative solution. The first item is again a majorizing

inequality. Note that the majorizing inequality is suggested by the algorithm, not the other way around. Define

$$\mathrm{BI}_r(x,y) \stackrel{\text{def}}{=} \omega\, L_r(x) + \sum_{p=1}^{\ell} (1-\omega\,\alpha_{rp}) \left\{ x_p \log \frac{x_p}{y_p} + y_p - x_p \right\}. \tag{268}$$

Lemma 24. *For nonnegative* $x,\ y \in \mathbb{R}^\ell$,

$$\mathrm{BI}_r(x,y) \leqslant \mathrm{BI}_r(y,y) + \mathrm{KL}(R_\omega\, y, x) - \mathrm{KL}(R_\omega\, y, y),$$

provided $\omega \leqslant 1/\max\{\alpha_{rp} : 1 \leqslant p \leqslant \ell\}$.

Proof. Apply Lemma 1 to $\omega\ \{\mathrm{KL}(B_r, A_r x) - \mathrm{KL}(B_r, A_r y)\}$. Also observe that

$$\sum_q [A_r x]_q = \sum_{p=1}^{\ell} \alpha_{rp}\, x_p,$$

and likewise for $[A_r y]_q$. This gives

$$\mathrm{BI}_r(x,y) \leqslant \mathrm{BI}_r(y,y) + \sum_{p=1}^{\ell} [R_\omega y]_p \log \frac{y_p}{x_p} + x_p - y_p,$$

and the lemma follows. Note that the condition on ω is used implicitly to assure that $R_{\omega,r}\, y$ is nonnegative. ■

The first monotonicity property is an easy consequence.

Lemma 25. *For* $r = k \bmod s$ *and* $\omega \leqslant 1/\max\{\alpha_{rp} : 1 \leqslant p \leqslant \ell\}$,

$$L_r(x^{[k]}) - L_r\left(x^{[k+1]}\right) \geqslant \omega^{-1}\, \mathrm{KL}\left(x^{[k+1]}, x^{[k]}\right) \geqslant 0.$$

Proof. Take $y = x^{[k]}$ and $x = R_\omega y = R_\omega x^{[k]} = x^{[k+1]}$. Then, one gets $\mathrm{BI}\left(x^{[k+1]}, x^{[k]}\right) - \mathrm{BI}\left(x^{[k]}, x^{[k]}\right) \leqslant -\mathrm{KL}\left(x^{[k+1]}, x^{[k]}\right)$, so that

$$\omega\, \left(L_r(x^{[k]}) - L_r\left(x^{[k+1]}\right)\right) \geqslant \sum_{p=1}^{\ell} (2-\omega\,\alpha_{rp}) \left\{ x_p \log \frac{x_p}{y_p} + y_p - x_p \right\}.$$

Since $\omega\,\alpha_{rp} \leqslant 1$, the conclusion follows. ■

The second monotonicity property reads as follows.

Lemma 26. *If $x^* \geqslant 0$ satisfies $Ax^* = b$, then with $r = k \bmod s$,*

$$\mathrm{KL}\left(x^*, x^{[k]}\right) - \mathrm{KL}\left(x^*, x^{[k+1]}\right) \geqslant \omega \left\{L_r\left(x^{[k]}\right) - L_r\left(x^*\right)\right\},$$

provided $\omega \leqslant 1/\max\{\alpha_{rp} : 1 \leqslant p \leqslant \ell\}$.

The lemma suggests that one should take ω as large as possible. This is how Byrne [9] arrived at the choice (267).

Proof of Lemma 26. Since x^* satisfies $Ax^* = b$, so $A_r x^* = B_r$ for all r, the proof is actually simpler than for the original proof for the EM algorithm; see section "Monotonicity of the Shepp–Vardi EM Algorithm." By the concavity of the logarithm (twice), one obtains

$$\begin{aligned}\log \frac{x_p^{[k+1]}}{x_p^{[k]}} &= \log\left((1 - \omega\alpha_{rp}) + \omega\left[A_r\left\{\frac{B_r}{A_r x^{[k]}}\right\}\right]_p\right) \\ &\geqslant \omega\,\alpha_{rp}\,\log\left(\alpha_{rp}^{-1}\left[A_r\left\{\frac{B_r}{A_r x^{[k]}}\right\}\right]_p\right) \geqslant \omega\left[A_r^T\,\log\left\{\frac{B_r}{A_r x^{[k]}}\right\}\right]_p,\end{aligned}$$

so that

$$\begin{aligned}\mathrm{KL}\left(x^*, x^{[k]}\right) - \mathrm{KL}\left(x^*, x^{[k+1]}\right) \geqslant \omega \sum_{p=1}^{\ell} x_p^*\left[A_r^T\,\log\left(B_r / A_r x^{[k]}\right)\right]_p \\ + \sum_{p=1}^{\ell} x_p^{[k]} - x_p^{[k+1]}.\end{aligned}$$

Now, the first sum equals

$$\sum_q [A_r x^*]_q\,\log\left([B_r]_q / [A_r x^{[k]}]_q\right) = \sum_q [B_r]_q\,\log\left([B_r]_q / [A_r x^{[k]}]_q\right).$$

For the remaining sums, note that

$$\begin{aligned}\sum_{p=1}^{\ell} x_p^{[k+1]} &= \sum_{p=1}^{\ell}\left(1 - \omega\,\alpha_{rp}\right) x_p^{[k]} + \omega\, x_p^{[k]}\left[A_r^T\left(B_r / A_r x^{[k]}\right)\right]_p \\ &= \sum_{p=1}^{\ell} x_p^{[k]} - \omega \sum_q \left[A_r x^{[k]}\right]_q + \omega \sum_q [B_r]_q.\end{aligned}$$

Putting the two together proves the lemma. ∎

Remark 5. To wrap things up, note that Byrne [9] shows the convergence (in the consistent case) of a somewhat different version of (266), which, under the subset-separability condition (229), reduces to the OSEM algorithm (224), thus proving the convergence of OSEM under subset-separability (in the consistent case). See also Corollary 1.

References

1. Aronszajn, N., Smith, K.T.: Theory of Bessel potentials. I. Ann. Inst. Fourier (Grenoble) **11**, 385–475 (1961). www.numdam.org
2. Atkinson, K.E.: The numerical solution of integral equations on the half line. SIAM J. Numer. Anal. **6**, 375–397 (1969)
3. Bardsley, J.M., Luttman, A.: Total variation-penalized Poisson likelihood estimation for ill-posed problems. Adv. Comput. Math. **31**, 35–39 (2009)
4. Bertero, M., Bocacci, P., Desiderá, G., Vicidomini, G.: Image de-blurring with Poisson data: from cells to galaxies. Inverse Probl. **25**(123006), 26 (2009)
5. Browne, J., De Pierro, A.R.: A row-action alternative to the EM algorithm for maximizing likelihoods in emission tomography. IEEE Trans. Med. Imaging **15**, 687–699 (1996)
6. Brune, C., Sawatzky, A., Burger, M.: Bregman-EM-TV methods with application to optical nanoscopy. In: Second International Conference on Scale Space and Variational Methods in Computer Vision, Voss. Lecture Notes in Computer Science, vol. 5567, pp. 235–246. Springer, Berlin (2009)
7. Byrne, C.L.: Iterative image reconstruction algorithms based on cross-entropy minimization. IEEE Trans. Image Process. **2**, 96–103 (1993)
8. Byrne, C.L.: Block-iterative methods for image reconstruction from projections. IEEE Trans. Image Process. **5**, 792–794 (1996)
9. Byrne, C.L.: Accelerating the EMML algorithm and related iterative algorithms by rescaled block-iterative methods. IEEE Trans. Image Process. **7**, 792–794 (1998)
10. Byrne, C.L.: Likelihood maximization for list-mode emission tomographic image reconstruction. IEEE Trans. Med. Imaging **20**, 1084–1092 (2001)
11. Byrne, C.L.: Choosing parameters in block-iterative or ordered subset reconstruction algorithms. IEEE Trans. Image Process. **14**, 321–327 (2005)
12. Byrne, C.L.: Signal Processing: A Mathematical Approach. AK Peters, Wellesley (2005)
13. Byrne, C.L.: Applied Iterative Methods. AK Peters, Wellesley (2008)
14. Byrne, C.L., Fiddy, M.A.: Images as power spectra; reconstruction as a Wiener filter approximation. Inverse Probl. **4**, 399–409 (1988)
15. Cao, Y.u., Eggermont, P.P.B., Terebey, S.: Cross Burg entropy maximization and its application to ringing suppression in image reconstruction. IEEE Trans. Image Process. **8**, 286–292 (1999)
16. Censor, Y., Eggermont, P.P.B., Gordon, D.: Strong under relaxation in Kaczmarz's method for inconsistent systems. Numer. Math. **41**, 83–92 (1983)
17. Censor, Y., Lent, A.H.: Optimization of "log x" entropy over linear equality constraints. SIAM J. Control. Optim. **25**, 921–933 (1987)
18. Censor, Y., Segman, J.: On block-iterative entropy maximization. J. Inf. Optim. Sci. **8**, 275–291 (1987)
19. Censor, Y., Zenios, S.A.: Proximal minimization algorithm with D-functions. J. Optim. Theory Appl. **73**, 451–464 (1992)
20. Cover, T.M.: An algorithm for maximizing expected log investment return. IEEE Trans. Inf. Theory **30**, 369–373 (1984)

21. Crowther, R.A., DeRosier, D.J., Klug, A.: The reconstruction of three-dimensional structure from projections and its application to electron microscopy. Proc. R. Soc. Lond. A Math. Phys. Sci. **317**(3), 19–340 (1971)
22. Csiszár, I.: I-divergence geometry of probability distributions and minimization problems. Ann. Probab. **3**, 146–158 (1975)
23. Csiszár, I., Tusnády, G.: Information geometry and alternating minimization procedures. Stat. Decis. **1**(Supplement 1), 205–237 (1984)
24. Daley, D.J., Vere-Jones, D.: An Introduction to the Theory of Point Processes. Springer, New York (2003)
25. Darroch, J.N., Ratcliff, D.: Generalized iterative scaling for log-linear models. Ann. Math. Stat. **43**, 1470–1480 (1972)
26. Daube-Witherspoon, M.E., Muehllehner, G.: An iterative space reconstruction algorithm suitable for volume ECT. IEEE Trans. Med. Imaging **5**, 61–66 (1986)
27. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. B **37**, 1–38 (1977)
28. De Pierro, A.R.: On the convergence of the iterative image space reconstruction algorithm for volume ECT. IEEE Trans. Med. Imaging **6**, 174–175 (1987)
29. De Pierro, A.R.: A modified expectation maximization algorithm for penalized likelihood estimation in emission tomography. IEEE Trans. Med. Imaging **14**, 132–137 (1995)
30. De Pierro, A., Yamaguchi, M.: Fast EM-like methods for maximum a posteriori estimates in emission tomography. Trans. Med. Imaging **20**, 280–288 (2001)
31. Dey, N., Blanc-Ferraud, L., Zimmer, Ch., Roux, P., Kam, Z., Olivo-Martin, J.-Ch., Zerubia, J.: Richardson-Lucy algorithm with total variation regularization for 3D confocal microscope deconvolution. Microsc. Res. Tech. **69**, 260–266 (2006)
32. Duijster, A., Scheunders, P., De Backer, S.: Wavelet-based EM algorithm for multispectral-image restoration. IEEE Trans. Geosci. Remote Sens. **47**, 3892–3898 (2009)
33. Eggermont, P.P.B.: Multiplicative iterative algorithms for convex programming. Linear Algebra Appl. **130**, 25–42 (1990)
34. Eggermont, P.P.B.: Nonlinear smoothing and the EM algorithm for positive integral equations of the first kind. Appl. Math. Optim. **39**, 75–91 (1999)
35. Eggermont, P.P.B., Herman, G.T., Lent, A.H.: Iterative algorithms for large partitioned linear systems with applications to image reconstruction. Linear Algebra Appl. **40**, 37–67 (1981)
36. Eggermont, P.P.B., LaRiccia, V.N.: Smoothed maximum likelihood density estimation for inverse problems. Ann. Stat. **23**, 199–220 (1995)
37. Eggermont, P.P.B., LaRiccia, V.N.: Maximum penalized likelihood estimation and smoothed EM algorithms for positive integral equations of the first kind. Numer. Funct. Anal. Optim. **17**, 737–754 (1997)
38. Eggermont, P.P.B., LaRiccia, V.N.: On EM-like algorithms for minimum distance estimation. Manuscript, University of Delaware (1998)
39. Eggermont, P.P.B., LaRiccia, V.N.: Maximum Penalized Likelihood Estimation, I: Density Estimation. Springer, New York (2001)
40. Elfving, T.: On some methods for entropy maximization and matrix scaling. Linear Algebra Appl. **34**, 321–339 (1980)
41. Fessler, J.A., Ficaro, E.P., Clinthorne, N.H., Lange, K.: Grouped coordinate ascent algorithms for penalized log-likelihood transmission image reconstruction. IEEE Trans. Med. Imaging **16**, 166–175 (1997)
42. Fessler, J.A., Hero, A.O.: Penalized maximum-likelihood image reconstruction using space-alternating generalized EM algorithms. IEEE Trans. Image Process. **4**, 1417–1429 (1995)
43. Figueiredo, M.A.T., Nowak, R.D.: An EM algorithm for wavelet-based image restoration. IEEE Trans. Image Process. **12**, 906–916 (2003)
44. Frank, J.: Three-Dimensional Electron Microscopy of Macromolecular Assemblies, 2nd edn. Oxford University Press, New York (2006)
45. Geman, S., Geman, D.: Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. IEEE Trans. Pattern Anal. Mach. Intell. **6**, 721–741 (1984)

46. Geman, S., McClure, D.E.: Bayesian image analysis, an application to single photon emission tomography. In: Proceedings of the Statistical Computing Section, Las Vegas, pp. 12–18. American Statistical Association (1985)
47. Good, I.J.: A nonparametric roughness penalty for probability densities. Nature **229**, 29–30 (1971)
48. Gordon, R., Bender, R., Herman, G.T.: Algebraic reconstruction techniques (ART) for three-dimensional electron microscopy and X-ray photography. J. Theor. Biol. **29**, 471–482 (1970)
49. Green, P.J.: Bayesian reconstructions from emission tomography data using a modified EM algorithm. IEEE Trans. Med. Imaging **9**, 84–93 (1990)
50. Guillaume, M., Melon, P., Réfrégier, P.: Maximum-likelihood estimation of an astronomical image from a sequence at low photon levels. J. Opt. Soc. Am. A **15**, 2841–2848 (1998)
51. Haltmeier, M., Leitão, A., Resmerita, E.: On regularization methods of EM-Kaczmarz type. Inverse Probl. **25**(075008), 17 (2009)
52. Hanke, M.: Accelerated Landweber iterations for the solution of ill-posed problems. Numer. Math. **60**, 341–373 (1991)
53. Hartley, H.O.: Maximum likelihood estimation from incomplete data. Biometrics **14**, 174–194 (1958)
54. Hebert, T., Leahy, R.: A generalized EM algorithm for 3-D Bayesian reconstruction from Poisson data using Gibbs priors. IEEE Trans. Med. Imaging **8**, 194–202 (1989)
55. Herman, G.T.: Fundamentals of Computerized Tomography: Image Reconstruction from Projections. Springer, New York (2009)
56. Herman, G.T., Meyer, L.B.: Algebraic reconstruction techniques can be made computationally efficient. IEEE Trans. Med. Imaging **12**, 600–609 (1993)
57. Holte, S., Schmidlin, P., Lindén, A., Rosenqvist, G., Eriksson, L.: Iterative image reconstruction for positron emission tomography: a study of convergence and quantitation problems. IEEE Trans. Nucl. Sci. **37**, 629–635 (1990)
58. Horváth, I., Bagoly, Z., Balász, L.G., de Ugarte Postigo, A., Veres, P., Mészáros, A.: Detailed classification of Swift's Gamma-ray bursts. J. Astrophys. **713**, 552–557 (2010)
59. Hudson, H.M., Larkin, R.S.: Accelerated image reconstruction using ordered subsets of projection data. IEEE Trans. Med. Imaging **13**, 601–609 (1994)
60. Kamphuis, C., Beekman, F.J., Viergever, M.A.: Evaluation of OS-EM vs. EM-ML for 1D, 2D and fully 3D SPECT reconstruction. IEEE Trans. Nucl. Sci. **43**, 2018–2024 (1996)
61. Kondor, A.: Method of convergent weights – an iterative procedure for solving Fredholm's integral equations of the first kind. Nucl. Instrum. Methods **216**, 177–181 (1983)
62. Lange, K.: Convergence of EM image reconstruction algorithms with Gibbs smoothing. IEEE Trans. Med. Imaging **9**, 439–446 (1990)
63. Lange, K., Bahn, M., Little, R.: A theoretical study of some maximum likelihood algorithms for emission and transmission tomography. IEEE Trans. Med. Imaging **6**, 106–114 (1987)
64. Lange, K., Carson, R.: EM reconstruction algorithms for emission and transmission tomography. J. Comput. Assist. Tomogr. **8**, 306–316 (1984)
65. Latham, G.A.: Existence of EMS solutions and a priori estimates. SIAM J. Matrix Anal. Appl. **16**, 943–953 (1995)
66. Levitan, E., Chan, M., Herman, G.T.: Image-modeling Gibbs priors. Graph. Models Image Process. **57**, 117–130 (1995)
67. Lewitt, R.M., Muehllehner, G.: Accelerated iterative reconstruction in PET and TOFPET. IEEE Trans. Med. Imaging **5**, 16–22 (1986)
68. Liu, C., Rubin, H.: The ECME algorithm: a simple extension of EM and ECM with faster monotone convergence. Biometrika **81**, 633–648 (1994)
69. Llacer, J., Veklerov, E.: Feasible images and practical stopping rules for iterative algorithms in emission tomography. IEEE Trans. Med. Imaging **8**, 186–193 (1989)
70. Lucy, L.B.: An iterative technique for the rectification of observed distributions. Astron. J. **79**, 745–754 (1974)
71. McLachlan, G.J., Krishnan, T.: The EM Algorithm and Its Extensions. Wiley, Hoboken (2008)

72. Meidunas, E.: Re-scaled block iterative expectation maximization maximum likelihood (RBI-EMML) abundance estimation and sub-pixel material identification in hyperspectral imagery. MS thesis, Department of Electrical Engineering, University of Massachusetts Lowell (2001)
73. Miller, M.I., Roysam, B.: Bayesian image reconstruction for emission tomography incorporating Good's roughness prior on massively parallel processors. Proc. Natl. Acad. Sci. U.S.A. **88**, 3223–3227 (1991)
74. Mülthei, H.N., Schorr, B.: On an iterative method for a class of integral equations of the first kind. Math. Methods Appl. Sci. **9**, 137–168 (1987)
75. Mülthei, H.N., Schorr, B.: On properties of the iterative maximum likelihood reconstruction method. Math. Methods Appl. Sci. **11**, 331–342 (1989)
76. Nielsen, S.F.: The stochastic EM algorithm: estimation and asymptotic results. Bernoulli **6**, 457–489 (2006)
77. Parra, L., Barrett, H.: List-mode likelihood: EM algorithm and image quality estimation demonstrated on 2-D PET. IEEE Trans. Med. Imaging **17**, 228–235 (1998)
78. Penczek, P., Zhu, J., Schroeder, R., Frank, J.: Three-dimensional reconstruction with contrast transfer function compensation. Scanning Microsc. **11**, 147–154 (1997)
79. Redner, R.A., Walker, H.F.: Mixture densities, maximum likelihood and the EM algorithm. SIAM Rev. **26**, 195–239 (1984)
80. Resmerita, E., Engl, H.W., Iusem, A.N.: The expectation-maximization algorithm for ill-posed integral equations: a convergence analysis. Inverse Probl. **23**, 2575–2588 (2007)
81. Richardson, W.H.: Bayesian based iterative method of image restoration. J. Opt. Soc. Am. **62**, 55–59 (1972)
82. Rockmore, A., Macovski, A.: A maximum likelihood approach to emission image reconstruction from projections. IEEE Trans. Nucl. Sci. **23**, 1428–1432 (1976)
83. Scheres, S.H.W., Gao, H.X., Valle, M., Herman, G.T., Eggermont, P.P.B., Frank, J., Carazo, J.-M.: Disentangling conformational states of macromolecules in 3D-EM through likelihood optimization. Nat. Methods **4**, 27–29 (2007)
84. Scheres, S.H.W., Núñez-Ramírez, R., Gómez-Llorente, Y., San Martín, C., Eggermont, P.P.B., Carazo, J.-M.: Modeling experimental image formation for likelihood-based classification of electron microscopy. Structure **15**, 1167–1177 (2007)
85. Scheres, S.H.W., Valle, M., Núñez, R., Sorzano, C.O.S., Marabini, R., Herman, G.T., Carazo, J.-M.: Maximum-likelihood multi-reference refinement for electron microscopy images. J. Mol. Biol. **348**, 139–149 (2005)
86. Schmidlin, P.: Iterative separation of tomographic scintigrams. Nuklearmedizin **11**, 1–16 (1972)
87. Setzer, S., Steidl, G., Teuber, T.: Deblurring Poissonian images by split Bregman techniques. J. Vis. Commun. Image Represent. **21**, 193–199 (2010)
88. Shepp, L.A., Vardi, Y.: Maximum likelihood reconstruction in emission tomography. IEEE Trans. Med. Imaging **1**, 113–122 (1982)
89. Sigworth, F.J.: A maximum-likelihood approach to single-particle image refinement. J. Struct. Biol. **122**, 328–339 (1998)
90. Silverman, B.W., Jones, M.C., Wilson, J.D., Nychka, D.W.: A smoothed EM algorithm approach to indirect estimation problems, with particular reference to stereology and emission tomography (with discussion). J. R. Stat. Soc. B **52**, 271–324 (1990)
91. Sun, Y., Walker, J.G.: Maximum likelihood data inversion for photon correlation spectroscopy. Meas. Sci. Technol. **19**(115302), 8 (2008)
92. Tanaka, E., Kudo, H.: Optimal relaxation parameters of DRAMA (dynamic RAMLA) aiming at one-pass image reconstruction for 3D-PET. Phys. Med. Biol. **55**, 2917–2939 (2010)
93. Tarasko, M.Z.: On a method for solution of the linear system with stochastic matrices (in Russian), Report Physics and Energetics Institute, Obninsk PEI-156 (1969)
94. Trummer, M.R.: A note on the ART of relaxation. Computing **33**, 349–352 (1984)
95. van der Sluis, A., van der Vorst, H.A.: SIRT- and CG-type methods for the iterative solution of sparse linear least-squares problems. Linear algebra in image reconstruction from projections. Linear Algebra Appl. **130**, 257–303 (1990)

96. Vardi, Y., Shepp, L.A., Kaufman, L.: A statistical model for positron emission tomography (with discussion). J. Am. Stat. Assoc. **80**, 8–38 (1985)
97. Wernick, M., Aarsvold, J.: Emission Tomography: The Fundamentals of PET and SPECT. Elsevier Academic, San Diego (2004)
98. Wu, C.F.J.: On the convergence properties of the EM algorithm. Ann. Stat. **11**, 95–103 (1983)
99. Yu, S., Latham, G.A., Anderssen, R.S.: Stabilizing properties of maximum penalized likelihood estimation for additive Poisson regression. Inverse Probl. **10**, 1199–1209 (1994)
100. Yuan, J., Yu, J.: Median-prior tomography reconstruction combined with nonlinear anisotropic diffusion filtering. J. Opt. Soc. Am. A **24**, 1026–1033 (2007)

EM Algorithms from a Non-stochastic Perspective

Charles Byrne

Contents

1 Introduction 390
2 A Non-stochastic Formulation of EM 391
The Non-stochastic EM Algorithm 391
3 The Stochastic EM Algorithm 393
The E-Step and M-Step 393
Difficulties with the Conventional Formulation 394
An Incorrect Proof 395
Acceptable Data 396
4 The Discrete Case 397
5 Missing Data 398
6 The Continuous Case 399
Acceptable Preferred Data 400
Selecting Preferred Data 401
Preferred Data as Missing Data 401
7 The Continuous Case with $Y = h(X)$ 402
An Example 402
Censored Exponential Data 403
A More General Approach 405
8 A Multinomial Example 406
9 The Example of Finite Mixtures 407
10 The EM and the Kullback-Leibler Distance 407
Using Acceptable Data 407
11 The Approach of Csiszár and Tusnády 409
The Framework of Csiszár and Tusnády 409
Alternating Minimization for the EM Algorithm 410
12 Sums of Independent Poisson Random Variables 412
Poisson Sums 412
The Multinomial Distribution 414

C. Byrne (✉)
Department of Mathematical Sciences, University of Massachusetts Lowell, Lowell, MA, USA
e-mail: charles_byrne@uml.edu

© Springer Science+Business Media New York 2015

O. Scherzer (ed.), *Handbook of Mathematical Methods in Imaging*,
DOI 10.1007/978-1-4939-0790-8_46

13 Poisson Sums in Emission Tomography 414
The SPECT Reconstruction Problem 415
Using the KL Distance 417
14 Nonnegative Solutions for Linear Equations 417
The General Case 418
Regularization 418
Acceleration 418
Using Prior Bounds on λ 420
15 Finite Mixture Problems 423
Mixtures 423
The Likelihood Function 423
A Motivating Illustration 424
The Acceptable Data 424
The Mix-EM Algorithm 425
Convergence of the Mix-EM Algorithm 426
16 More on Convergence 426
17 Open Questions 427
18 Conclusion 427
Cross-References 428
References 428

Abstract

The EM algorithm is not a single algorithm, but a template for the construction of iterative algorithms. While it is always presented in stochastic language, relying on conditional expectations to obtain a method for estimating parameters in statistics, the essence of the EM algorithm is not stochastic. The conventional formulation of the EM algorithm given in many texts and papers on the subject is inadequate. A new formulation is given here based on the notion of acceptable data.

1 Introduction

The "expectation maximization" (EM) algorithm is a general framework for maximizing the likelihood function in statistical parameter estimation [1–3]. It is always presented in probabilistic terms, involving the maximization of a conditional expected value. The EM algorithm is not really a single algorithm, but a framework for the design of iterative likelihood maximization methods, or, as the authors of [4] put it, a "prescription for constructing an algorithm"; nevertheless, we shall continue to refer to the EM algorithm. As we shall demonstrate in Sect. 2, the essence of the EM algorithm is not stochastic. Our non-stochastic EM (NSEM) is a general approach for function maximization that has the stochastic EM methods as particular cases.

Maximizing the likelihood function is a well-studied procedure for estimating parameters from observed data. When a maximizer cannot be obtained in closed form, iterative maximization algorithms, such as the expectation maximization (EM) maximum likelihood algorithms, are needed. The standard formulation of the EM algorithms postulates that finding a maximizer of the likelihood is

complicated because the observed data is somehow incomplete or deficient, and the maximization would have been simpler had we observed the complete data. The EM algorithm involves repeated calculations involving complete data that has been estimated using the current parameter value and conditional expectation.

The standard formulation is adequate for the most common discrete case, in which the random variables involved are governed by finite or infinite probability functions, but unsatisfactory in general, particularly in the continuous case, in which probability density functions and integrals are needed.

We adopt the view that the observed data is not necessarily incomplete, but just difficult to work with, while different data, which we call the preferred data, leads to simpler calculations. To relate the preferred data to the observed data, we assume that the preferred data is *acceptable*, which means that the conditional distribution of the preferred data, given the observed data, is independent of the parameter. This extension of the EM algorithms contains the usual formulation for the discrete case, while removing the difficulties associated with the continuous case. Examples are given to illustrate this new approach.

2 A Non-stochastic Formulation of EM

The essence of the EM algorithm is not stochastic and leads to a general approach for function maximization, which we call the "non-stochastic," EM algorithm (NSEM) [6]. In addition to being more general, this new approach also simplifies much of the development of the EM algorithm itself.

The Non-stochastic EM Algorithm

We present now the essential aspects of the EM algorithm without relying on statistical concepts. We shall use these results later to establish important facts about the statistical EM algorithm. For a broader treatment of the EM algorithm in the context of iterative optimization, see [5].

The Continuous Case

The problem is to maximize a nonnegative function $f : Z \to \mathbb{R}$, where Z is an arbitrary set. We assume that there is $z^* \in Z$ with $f(z^*) \geq f(z)$, for all $z \in Z$. We also assume that there is a nonnegative function $b : \mathbb{R}^N \times Z \to \mathbb{R}$ such that

$$f(z) = \int b(x, z)dx.$$

Having found z^k, we maximize the function

$$H(z^k, z) = \int b(x, z^k) \log b(x, z)dx \tag{1}$$

to get z^{k+1}. Adopting such an iterative approach presupposes that maximizing $H(z^k, z)$ is simpler than maximizing $f(z)$ itself. This is the case with the EM algorithm.

The cross-entropy or Kullback-Leibler distance [7] is a useful tool for analyzing the EM algorithm. For positive numbers u and v, the Kullback-Leibler distance from u to v is

$$KL(u, v) = u \log \frac{u}{v} + v - u. \tag{2}$$

We also define $KL(0, 0) = 0$, $KL(0, v) = v$, and $KL(u, 0) = +\infty$. The KL distance is extended to nonnegative vectors component-wise, so that for nonnegative vectors a and b, we have

$$KL(a, b) = \sum_{j=1}^{J} KL(a_j, b_j). \tag{3}$$

One of the most useful and easily proved facts about the KL distance is contained in the following lemma; we simplify the notation by setting $b(z) = b(x, z)$.

Lemma 1. *For nonnegative vectors a and b, with $b_+ = \sum_{j=1}^{J} b_j > 0$, we have*

$$KL(a, b) = KL(a_+, b_+) + KL(a, \frac{a_+}{b_+} b). \tag{4}$$

This lemma can be extended to obtain the following useful identity.

Lemma 2. *For $f(z)$ and $b(x, z)$ as above, and z and w in Z, with $f(w) > 0$, we have*

$$KL(b(z), b(w)) = KL(f(z), f(w)) + KL(b(z), (f(z)/f(w))b(w)). \tag{5}$$

Maximizing $H(z^k, z)$ is equivalent to minimizing

$$G(z^k, z) = KL(b(z^k), b(z)) - f(z), \tag{6}$$

where

$$KL(b(z^k), b(z)) = \int KL(b(x, z^k), b(x, z))dx. \tag{7}$$

Therefore,

$$-f(z^k) = KL(b(z^k), b(z^k)) - f(z^k) \geq KL(b(z^k), b(z^{k+1})) - f(z^{k+1}),$$

or

$$f(z^{k+1}) - f(z^k) \geq KL(b(z^k), b(z^{k+1})) \geq KL(f(z^k), f(z^{k+1})).$$

Consequently, the sequence $\{f(z^k)\}$ is increasing and bounded above, so that the sequence $\{KL(b(z^k), b(z^{k+1}))\}$ converges to zero. Without additional restrictions, we cannot conclude that $\{f(z^k)\}$ converges to $f(z^*)$.

We get z^{k+1} by minimizing $G(z^k, z)$. When we minimize $G(z, z^{k+1})$, we get z^{k+1} again. Therefore, we can put the NSEM algorithm into the alternating-minimization (AM) framework of Csiszár and Tusnády [12], to be discussed further Sect. 11.

The Discrete Case

Again, the problem is to maximize a nonnegative function $f : Z \to \mathbb{R}$, where Z is an arbitrary set. As previously, we assume that there is $z^* \in Z$ with $f(z^*) \geq f(z)$, for all $z \in Z$. We also assume that there is a finite or countably infinite set B and a nonnegative function $b : B \times Z \to \mathbb{R}$ such that

$$f(z) = \sum_{x \in B} b(x, z).$$

Having found z^k, we maximize the function

$$H(z^k, z) = \sum_{x \in B} b(x, z^k) \log b(x, z) \tag{8}$$

to get z^{k+1}.

We set $b(z) = b(x, z)$ again. Maximizing $H(z^k, z)$ is equivalent to minimizing

$$G(z^k, z) = KL(b(z^k), b(z)) - f(z), \tag{9}$$

where

$$KL(b(z^k), b(z)) = \sum_{x \in B} KL(b(x, z^k), b(x, z)). \tag{10}$$

As previously, we find that the sequence $\{f(z^k)\}$ is increasing, and $\{KL(b(z^k), b(z^{k+1}))\}$ converges to zero. Without additional restrictions, we cannot conclude that $\{f(z^k)\}$ converges to $f(z^*)$.

3 The Stochastic EM Algorithm

The E-Step and M-Step

In statistical parameter estimation, one typically has an *observable* random vector Y taking values in $\mathbb{R}^N$ that is governed by a probability density function (pdf) or probability function (pf) of the form $f_Y(y|\theta)$, for some value of the parameter vector

$\theta \in \Theta$, where Θ is the set of all legitimate values of θ. Our *observed* data consists of one realization y of Y; we do not exclude the possibility that the entries of y are independently obtained samples of a common real-valued random variable. The true vector of parameters is to be estimated by maximizing the likelihood function $L_y(\theta) = f_Y(y|\theta)$ over all $\theta \in \Theta$ to obtain a maximum likelihood estimate, θ_{ML}.

To employ the EM algorithmic approach, it is assumed that there is another related random vector X, which we shall call the *preferred* data, such that, had we been able to obtain one realization x of X, maximizing the likelihood function $L_x(\theta) = f_X(x|\theta)$ would have been simpler than maximizing the likelihood function $L_y(\theta) = f_Y(y|\theta)$. Of course, we do not have a realization x of X. The basic idea of the EM approach is to estimate x using the current estimate of θ, denoted θ^k, and to use each estimate x^k of x to get the next estimate θ^{k+1}.

The EM algorithm proceeds in two steps. Having selected the preferred data X, and having found θ^k, we form the function of θ given by

$$Q(\theta|\theta^k) = E(\log f_X(x|\theta)|y, \theta^k); \tag{11}$$

this is the E-step of the EM algorithm. Then we maximize $Q(\theta|\theta^k)$ over all θ to get θ^{k+1}; this is the M-step of the EM algorithm. In this way, the EM algorithm based on X generates a sequence $\{\theta^k\}$ of parameter vectors.

For the discrete case of probability functions, we have

$$Q(\theta|\theta^k) = \sum_x f_{X|Y}(x|y, \theta^k) \log f_X(x|\theta), \tag{12}$$

and for the continuous case of probability density functions, we have

$$Q(\theta|\theta^k) = \int f_{X|Y}(x|y, \theta^k) \log f_X(x|\theta) dx. \tag{13}$$

In decreasing order of importance and difficulty, the goals are these:

1. To have the sequence of parameters $\{\theta^k\}$ converging to θ_{ML};
2. To have the sequence of functions $\{f_X(x|\theta^k)\}$ converging to $f_X(x|\theta_{ML})$;
3. To have the sequence of numbers $\{L_y(\theta^k)\}$ converging to $L_y(\theta_{ML})$;
4. To have the sequence of numbers $\{L_y(\theta^k)\}$ non-decreasing.

Our focus here is mainly on the fourth goal, with some discussion of the third goal. We do present some examples for which all four goals are attained. Clearly, the first goal requires a topology on the set Θ.

Difficulties with the Conventional Formulation

In [1] we are told that

$$f_{X|Y}(x|y,\theta) = f_X(x|\theta)/f_Y(y|\theta). \tag{14}$$

This is false; integrating with respect to x gives one on the left side and $1/f_Y(y|\theta)$ on the right side. Perhaps the equation is not meant to hold for all x, but just for some x. In fact, if there is a function h such that $Y = h(X)$, then Eq. (14) might hold for those x such that $h(x) = y$. In fact, this is what happens in the discrete case of probabilities; in that case we do have

$$f_Y(y|\theta) = \sum_{x \in h^{-1}\{y\}} f_X(x|\theta), \tag{15}$$

where

$$h^{-1}\{y\} = \{x|h(x) = y\}.$$

Consequently,

$$f_{X|Y}(x|y,\theta) = \begin{cases} f_X(x|\theta)/f_Y(y|\theta), & \text{if } x \in h^{-1}\{y\}; \\ 0, & \text{if } x \notin h^{-1}\{y\}. \end{cases} \tag{16}$$

However, this modification of Eq. (14) fails in the continuous case of probability density functions, since $h^{-1}\{y\}$ is often a subset of zero measure. Even if the set $h^{-1}\{y\}$ has positive measure, integrating both sides of Eq. (14) over $x \in h^{-1}\{y\}$ tells us that $f_Y(y|\theta) \leq 1$, which need not hold for probability density functions.

An Incorrect Proof

Everyone who works with the EM algorithm will say that the likelihood is non-decreasing for the EM algorithm. The proof of this fact usually proceeds as follows; we use the notation for the continuous case, but the proof for the discrete case is essentially the same. Use Eq. (14) to get

$$\log f_X(x|\theta) = \log f_{X|Y}(x|y,\theta) - \log f_Y(y|\theta). \tag{17}$$

Then replace the term $\log f_X(x|\theta)$ in Eq. (13) with the right side of Eq. (17), obtaining

$$\log f_Y(y|\theta) - Q(\theta|\theta^k) = -\int f_{X|Y}(x|y,\theta^k) \log f_{X|Y}(x|y,\theta) dx. \tag{18}$$

Jensen's Inequality tells us that

$$\int u(x) \log u(x) dx \geq \int u(x) \log v(x) dx, \tag{19}$$

for any probability density functions $u(x)$ and $v(x)$. Since $f_{X|Y}(x|y,\theta)$ is a probability density function, we have

$$\int f_{X|Y}(x|y,\theta^k)\log f_{X|Y}(x|y,\theta)dx \leq \int f_{X|Y}(x|y,\theta^k)\log f_{X|Y}(x|y,\theta^k)dx. \tag{20}$$

We conclude, therefore, that $\log f_Y(y|\theta) - Q(\theta|\theta^k)$ attains its minimum value at $\theta = \theta^k$. Then we have

$$\log f_Y(y|\theta^{k+1}) - \log f_Y(y|\theta^k) \geq Q(\theta^{k+1}|\theta^k) - Q(\theta^k|\theta^k) \geq 0. \tag{21}$$

This proof is incorrect; clearly it rests on the validity of Eq. (14), which is generally false. For the discrete case, with $Y = h(X)$, this proof is valid, when we use Eq. (16), instead of Eq. (14). In all other cases, however, the proof is incorrect.

Acceptable Data

We turn now to the question of how to repair the incorrect proof. Equation (14) should read

$$f_{X|Y}(x|y,\theta) = f_{X,Y}(x,y|\theta)/f_Y(y|\theta), \tag{22}$$

for all x. In order to replace $\log f_X(x|\theta)$ in Eq. (13), we write

$$f_{X,Y}(x,y|\theta) = f_{X|Y}(x|y,\theta)f_Y(y|\theta), \tag{23}$$

and

$$f_{X,Y}(x,y|\theta) = f_{Y|X}(y|x,\theta)f_X(x|\theta), \tag{24}$$

so that

$$\log f_X(x|\theta) = \log f_{X|Y}(x|y,\theta) + \log f_Y(y|\theta) - \log f_{Y|X}(y|x,\theta). \tag{25}$$

We say that the preferred data is *acceptable* if

$$f_{Y|X}(y|x,\theta) = f_{Y|X}(y|x); \tag{26}$$

that is, the dependence of Y on X is unrelated to the value of the parameter θ. This definition provides our generalization of the relationship $Y = h(X)$.

When X is acceptable, we have that $\log f_Y(y|\theta) - Q(\theta|\theta^k)$ again attains its minimum value at $\theta = \theta^k$. The assertion that the likelihood is non-decreasing then follows, using the same argument as in the previous incorrect proof.

4 The Discrete Case

In the discrete case, we assume that Y is a discrete random vector taking values in a finite or countably infinite set A, and governed by probability $f_Y(y|\theta)$. We assume, in addition, that there is a second discrete random vector X, taking values in a finite or countably infinite set B, and a function $h : B \to A$ such that $Y = h(X)$. We define the set

$$h^{-1}\{y\} = \{x \in B | h(x) = y\}. \tag{27}$$

Then we have

$$f_Y(y|\theta) = \sum_{x \in h^{-1}\{y\}} f_X(x|\theta). \tag{28}$$

The conditional probability function for X, given $Y = y$, is

$$f_{X|Y}(x|y, \theta) = \frac{f_X(x|\theta)}{f_Y(y|\theta)}, \tag{29}$$

for $x \in h^{-1}\{y\}$, and zero, otherwise. The so-called E-step of the EM algorithm is then to calculate

$$Q(\theta|\theta^k) = E((\log f_X(X|\theta)|y, \theta^k) = \sum_{x \in h^{-1}\{y\}} f_{X|Y}(x|y, \theta^k) \log f_X(x|\theta), \tag{30}$$

and the M-step is to maximize $Q(\theta|\theta^k)$ as a function of θ to obtain θ^{k+1}.

Using Eq. (29), we can write

$$Q(\theta|\theta^k) = \sum_{x \in h^{-1}\{y\}} f_{X|Y}(x|y, \theta^k) \log f_{X|Y}(x|y, \theta) + \log f_Y(y|\theta). \tag{31}$$

Therefore,

$$\log f_Y(y|\theta) - Q(\theta|\theta^k) = -\sum_{x \in h^{-1}\{y\}} f_{X|Y}(x|y, \theta^k) \log f_{X|Y}(x|y, \theta).$$

Since

$$\sum_{x \in h^{-1}\{y\}} f_{X|Y}(x|y, \theta^k) = \sum_{x \in h^{-1}\{y\}} f_{X|Y}(x|y, \theta) = 1,$$

it follows from Jensen's Inequality that

$$-\sum_{x\in h^{-1}\{y\}} f_{X|Y}(x|y,\theta^k)\log f_{X|Y}(x|y,\theta)$$
$$\geq -\sum_{x\in h^{-1}\{y\}} f_{X|Y}(x|y,\theta^k)\log f_{X|Y}(x|y,\theta^k).$$

Therefore, $\log f_Y(y|\theta) - Q(\theta|\theta^k)$ attains its minimum at $\theta = \theta^k$. We have the following result.

Proposition 1. *The sequence $\{f_Y(y|\theta^k)\}$ is non-decreasing.*

Proof. We have

$$\log f_Y(y|\theta^{k+1}) - Q(\theta^{k+1}|\theta^k) \geq \log f_Y(y|\theta^k) - Q(\theta^k|\theta^k),$$

or

$$\log f_Y(y|\theta^{k+1}) - \log f_Y(y|\theta^k) \geq Q(\theta^{k+1}|\theta^k) - Q(\theta^k|\theta^k) \geq 0.$$

■

Let $\chi_{h^{-1}\{y\}}(x)$ be the characteristic function of the set $h^{-1}\{y\}$, that is,

$$\chi_{h^{-1}\{y\}}(x) = \begin{cases} 1, \text{ if } x \in h^{-1}\{y\}; \\ 0, \text{ if } x \notin h^{-1}\{y\}. \end{cases} \tag{32}$$

With the choices $z = \theta$, $f(z) = f_Y(y|\theta)$, and $b(z) = f_X(x|\theta)\chi_{h^{-1}\{y\}}(x)$, the discrete EM algorithm fits into the framework of the non-stochastic EM algorithm. Consequently, we see once again that the sequence $\{f_Y(y|\theta^k)\}$ is non-decreasing and also that the sequence

$$\{KL(b(z^k), b(z^{k+1})\} = \{\sum_{x\in h^{-1}\{y\}} (KL(f_X(x|\theta^k), f_X(x|\theta^{k+1}))\}$$

converges to zero.

5 Missing Data

We say that there is *missing data* if the preferred data X has the form $X = (Y, W)$, so that $Y = h(X) = h(Y, W)$, where h is the orthogonal projection onto the first component. The case of missing data for the discrete case is covered by the discussion in Sect. 4, so we consider here the continuous case in which probability density functions are involved.

Once again, the E-step is to calculate $Q(\theta|\theta^k)$ given by

$$Q(\theta|\theta^k) = E(\log f_X(X|\theta)|y, \theta^k). \tag{33}$$

Since $X = (Y, W)$, we have

$$f_X(x|\theta) = f_{Y,W}(y, w|\theta). \tag{34}$$

Since the set $h^{-1}\{y\}$ has measure zero, we cannot write

$$Q(\theta|\theta^k) = \int_{h^{-1}\{y\}} f_{X|Y}(x|y, \theta^k) \log f_X(x|\theta) dx.$$

Instead, following [8], we write

$$Q(\theta|\theta^k) = \int f_{Y,W}(y, w|\theta^k) \log f_{Y,W}(y, w|\theta) dw / f_Y(y|\theta^k). \tag{35}$$

Consequently, maximizing $Q(\theta|\theta^k)$ is equivalent to maximizing

$$\int f_{Y,W}(y, w|\theta^k) \log f_{Y,W}(y, w|\theta) dw.$$

With $b(\theta) = b(\theta, w) = f_{Y,W}(y, w|\theta)$ and

$$f_Y(y|\theta) = f(\theta) = \int f_{Y,W}(y, w|\theta) dw = \int b(\theta) dw,$$

we find that maximizing $Q(\theta|\theta^k)$ is equivalent to minimizing $KL(b(\theta^k), b(\theta)) - f(\theta)$. Therefore, the EM algorithm for the case of missing data falls into the framework of the non-stochastic EM algorithm. We conclude that the sequence $\{f(\theta^k)\}$ is non-decreasing and that the sequence $\{KL(b(\theta^k), b(\theta^{k+1}))\}$ converges to zero.

Most other instances of the continuous case in which we have $Y = h(X)$ can be handled using the missing-data model. For example, suppose that Z_1 and Z_2 are uniformly distributed on the interval $[0, \theta]$, for some positive θ and that $Y = Z_1 + Z_2$. We may, for example, then take W to be $W = Z_1 - Z_2$ and $X = (Y, W)$ as the preferred data. We shall discuss these instances further in Sect. 7.

6 The Continuous Case

We turn now to the general continuous case. We have a random vector Y taking values in $\mathbb{R}^N$ and governed by the probability density function $f_Y(y|\theta)$. The objective, once again, is to maximize the likelihood function $L_y(\theta) = f_Y(y|\theta)$ to obtain the maximum likelihood estimate of θ.

Acceptable Preferred Data

For the continuous case, the vector θ^{k+1} is obtained from θ^k by maximizing the conditional expected value

$$Q(\theta|\theta^k) = E(\log f_X(X|\theta)|y, \theta^k) = \int f_{X|Y}(x|y, \theta^k) \log f_X(x|\theta) dx. \quad (36)$$

Assuming the acceptability condition and using

$$f_{X,Y}(x, y|\theta^k) = f_{X|Y}(x|y, \theta^k) f_Y(y|\theta^k),$$

and

$$\log f_X(x|\theta) = \log f_{X,Y}(x, y|\theta) - \log f_{Y|X}(y|x),$$

we find that maximizing $E(\log f_X(x|\theta)|y, \theta^k)$ is equivalent to minimizing

$$H(\theta^k, \theta) = \int f_{X,Y}(x, y|\theta^k) \log f_{X,Y}(x, y|\theta) dx. \quad (37)$$

With $f(\theta) = f_Y(y|\theta)$, and $b(\theta) = f_{X,Y}(x, y|\theta)$, this problem fits the framework of the non-stochastic EM algorithm and is equivalent to minimizing

$$G(\theta^k, \theta) = KL(b(\theta^k), b(\theta)) - f(\theta).$$

Once again, we may conclude that the likelihood function is non-decreasing and that the sequence $\{KL(b(\theta^k), b(\theta^{k+1}))\}$ converges to zero.

In the discrete case in which $Y = h(X)$, the conditional probability $f_{Y|X}(y|x, \theta)$ is $\delta(y - h(x))$, as a function of y, for given x, and is the characteristic function of the set $h^{-1}(y)$, as a function of x, for given y. Therefore, we can write $f_{X|Y}(x|y, \theta)$ using Eq. (16). For the continuous case in which $Y = h(X)$, the pdf $f_{Y|X}(y|x, \theta)$ is again a delta function of y, for given x; the difficulty arises when we need to view this as a function of x, for given y. The acceptability property helps us avoid this difficulty.

When X is acceptable, we have

$$f_{X|Y}(x|y, \theta) = f_{Y|X}(y|x) f_X(x|\theta)/f_Y(y|\theta), \quad (38)$$

whenever $f_Y(y|\theta) \neq 0$, and is zero otherwise. Consequently, when X is acceptable, we have a kernel model for $f_Y(y|\theta)$ in terms of the $f_X(x|\theta)$:

$$f_Y(y|\theta) = \int f_{Y|X}(y|x) f_X(x|\theta) dx; \quad (39)$$

for the continuous case we view this as a corrected version of Eq. (15). In the discrete case the integral is replaced by a summation, of course, but when we are speaking generally about either case, we shall use the integral sign.

The acceptability of the missing data W is used in [9], but more for computational convenience and to involve the Kullback-Leibler distance in the formulation of the EM algorithm. It is not necessary that W be acceptable in order for likelihood to be non-decreasing, as we have seen.

Selecting Preferred Data

The popular example of multinomial data given below illustrates well the point that one can often choose to view the observed data as "incomplete" simply in order to introduce "complete" data that makes the calculations simpler, even when there is no suggestion, in the original problem, that the observed data is in any way inadequate or "incomplete." It is in order to emphasize this desire for simplification that we refer to X as the preferred data, not the complete data.

In some applications, the preferred data X arises naturally from the problem, while in other cases the user must imagine preferred data. This choice in selecting the preferred data can be helpful in speeding up the algorithm (see [10]).

If, instead of maximizing

$$\int f_{X|Y}(x|y,\theta^k)\log f_X(x|\theta)dx,$$

at each M-step, we simply select θ^{k+1} so that

$$\int f_{X|Y}(x|y,\theta^k)\log f_{X,Y}(x,y|\theta^{k+1})dx - \int f_{X|Y}(x|y,\theta^k)\log f_{X,Y}(x,y|\theta^k)dx > 0,$$

we say that we are using a *generalized* EM (GEM) algorithm. It is clear from the discussion in the previous subsection that whenever X is acceptable, a GEM also guarantees that likelihood is non-decreasing.

Preferred Data as Missing Data

As we have seen, when the EM algorithm is applied to the missing-data model, the likelihood is non-decreasing, which suggests that, for an arbitrary preferred data X, we could imagine X as W, the missing data, and imagine applying the EM algorithm to $Z = (Y, X)$. This approach would produce an EM sequence of parameter vectors for which likelihood is non-decreasing, but it need not be the

same sequence as obtained by applying the EM algorithm to X directly. It is the same sequence, provided that X is acceptable. We are not suggesting that applying the EM algorithm to $Z = (Y, X)$ would simplify calculations.

We know that, when the missing-data model is used and the M-step is defined as maximizing the function in (35), the likelihood is not decreasing. It would seem then that for any choice of preferred data X, we could view this data as missing and take as our complete data the pair $Z = (Y, X)$, with X now playing the role of W. Maximizing the function in (35) is then equivalent to maximizing

$$\int f_{X|Y}(x|y, \theta^k) \log f_{X,Y}(x, y|\theta) dx; \tag{40}$$

to get θ^{k+1}. It then follows that $L_y(\theta^{k+1}) \geq L_y(\theta^k)$. The obvious question is whether or not these two functions given in (11) and (40) have the same maximizers.

For acceptable X we have

$$\log f_{X,Y}(x, y|\theta) = \log f_X(x|\theta) + \log f_{Y|X}(y|x), \tag{41}$$

so the two functions given in (11) and (40) do have the same maximizers. It follows once again that whenever the preferred data is acceptable, we have $L_y(\theta^{k+1}) \geq L_y(\theta^k)$. Without additional assumptions, however, we cannot conclude that $\{\theta^k\}$ converges to θ_{ML}, nor that $\{f_Y(y|\theta^k)\}$ converges to $f_Y(y|\theta_{ML})$.

7 The Continuous Case with $Y = h(X)$

In this section we consider the continuous case in which the observed random vector Y takes values in $\mathbb{R}^N$; the preferred random vector X takes values in $\mathbb{R}^M$; the random vectors are governed by probability density functions $f_Y(y|\theta)$ and $f_X(x|\theta)$, respectively; and there is a function $h : \mathbb{R}^N \to \mathbb{R}^M$ such that $Y = h(X)$. In most cases, $M > N$ and $h^{-1}\{y\} = \{x|h(x) = y\}$ has measure zero in $\mathbb{R}^M$.

An Example

For example, suppose that Z_1 and Z_2 are independent and uniformly distributed on the interval $[0, \theta]$, for some $\theta > 0$ to be estimated. Let $Y = Z_1 + Z_2$. With $Z = (Z_1, Z_2)$, and $h : \mathbb{R}^2 \to \mathbb{R}$ given by $h(z_1, z_2) = z_1 + z_2$, we have $Y = h(Z)$. The pdf for Z is

$$f_Z(z|\theta) = f_Z(z_1, z_2|\theta) = \frac{1}{\theta^2} \chi_{[0,\theta]}(z_1) \chi_{[0,\theta]}(z_2). \tag{42}$$

The pdf for Y is

$$f_Y(y|\theta) = \begin{cases} \frac{y}{\theta^2}, & \text{if } 0 \le y \le \theta; \\ \frac{2\theta - y}{\theta^2}, & \text{if } \theta \le y \le 2\theta. \end{cases} \tag{43}$$

It is not the case that

$$f_Y(y|\theta) = \int_{h^{-1}\{y\}} f_Z(z|\theta), \tag{44}$$

since $h^{-1}\{y\}$ has measure zero in $\mathbb{R}^2$.

The likelihood function is $L(\theta) = f_Y(y|\theta)$, viewed as a function of θ, and is given by

$$L(\theta) = \begin{cases} \frac{y}{\theta^2}, & \text{if } \theta \ge y; \\ \frac{2\theta - y}{\theta^2}, & \text{if } \frac{y}{2} \le \theta \le y. \end{cases} \tag{45}$$

Therefore, the maximum likelihood estimate of θ is $\theta_{ML} = y$.

Instead of using Z as our preferred data, suppose that we define the random variable $W = Z_2$, and let $X = (Y, W)$, a missing-data model. We then have $Y = h(X)$, where $h : \mathbb{R}^2 \to \mathbb{R}$ is given by $h(x) = h(y, w) = y$. The pdf for Y given in Eq. (43) can be written as

$$f_Y(y|\theta) = \int \frac{1}{\theta^2} \chi_{[0,\theta]}(y - w) \chi_{[0,\theta]}(w) dw. \tag{46}$$

The joint pdf is

$$f_{Y,W}(y, w|\theta) = \begin{cases} 1/\theta^2, & \text{for } w \le y \le \theta + w; \\ 0, & \text{otherwise.} \end{cases} \tag{47}$$

Censored Exponential Data

McLachlan and Krishnan [1] give the following example of a likelihood maximization problem involving probability density functions. This example provides a good illustration of the usefulness of the missing-data model.

Suppose that Z is the time until failure of a component, which we assume is governed by the exponential distribution

$$f(z|\theta) = \frac{1}{\theta} e^{-z/\theta}, \tag{48}$$

where the parameter $\theta > 0$ is the expected time until failure. We observe a random sample of N components and record their failure times, z_n. On the basis of this data, we must estimate θ, the mean time until failure.

It may well happen, however, that during the time allotted for observing the components, only r of the N components fail, which, for convenience, are taken to be the first r items in the record. Rather than wait longer, we record the failure times of those that failed and record the elapsed time for the experiment, say T, for those that had not yet failed. The *censored data* is then $y = (y_1, \ldots, y_N)$, where $y_n = z_n$ is the time until failure for $n = 1, \ldots, r$, and $y_n = T$ for $n = r+1, \ldots, N$. The censored data is reasonably viewed as *incomplete*, relative to the *complete* data we would have had, had the trial lasted until all the components had failed.

Since the probability that a component will survive until time T is $e^{-T/\theta}$, the pdf for the vector y is

$$f_Y(y|\theta) = \Big(\prod_{n=1}^{r} \frac{1}{\theta} e^{-y_n/\theta}\Big)\, e^{-(N-r)T/\theta}, \tag{49}$$

and the log likelihood function for the censored, or incomplete, data is

$$\log f_Y(y|\theta) = -r \log \theta - \frac{1}{\theta} \sum_{n=1}^{N} y_n. \tag{50}$$

In this particular example, we are fortunate, in that we can maximize $f_Y(y|\theta)$ easily, and find that the ML solution based on the incomplete, censored data is

$$\theta_{MLi} = \frac{1}{r} \sum_{n=1}^{N} y_n = \frac{1}{r} \sum_{n=1}^{r} y_n + \frac{N-r}{r} T. \tag{51}$$

In most cases in which our data is incomplete, finding the ML estimate from the incomplete data is difficult, while finding it for the complete data is relatively easy.

We say that the missing data are the times until failure of those components that did not fail during the observation time. The preferred data is the complete data $x = (z_1, \ldots, z_N)$ of actual times until failure. The pdf for the preferred data X is

$$f_X(x|\theta) = \prod_{n=1}^{N} \frac{1}{\theta} e^{-z_n/\theta}, \tag{52}$$

and the log likelihood function based on the complete data is

$$\log f_X(x|\theta) = -N \log \theta - \frac{1}{\theta} \sum_{n=1}^{N} z_n. \tag{53}$$

The ML estimate of θ from the complete data is easily seen to be

$$\theta_{MLc} = \frac{1}{N} \sum_{n=1}^{N} z_n. \tag{54}$$

In this example, both the incomplete-data vector y and the preferred-data vector x lie in $\mathbb{R}^N$. We have $y = h(x)$ where the function h operates by setting to T any component of x that exceeds T. Clearly, for a given y, the set $h^{-1}\{y\}$ consists of all vectors x with entries $x_n \geq T$ or $x_n = y_n < T$. For example, suppose that $N = 2$, and $y = (y_1, T)$, where $y_1 < T$. Then $h^{-1}\{y\}$ is the one-dimensional ray

$$h^{-1}\{y\} = \{x = (y_1, x_2) | \, x_2 \geq T\}.$$

Because this set has measure zero in $\mathbb{R}^2$, Eq. (44) does not make sense in this case.

We need to calculate $E(\log f_X(X|\theta)|y, \theta^k)$. Following McLachlan and Krishnan [1], we note that since $\log f_X(x|\theta)$ is linear in the unobserved data Z_n, $n = r + 1, \ldots, N$, to calculate $E(\log f_X(X|\theta)|y, \theta^k)$, we need only replace the unobserved values with their conditional expected values, given y and θ^k. The conditional distribution of $Z_n - T$, given that $Z_n > T$, is still exponential, with mean θ. Therefore, we replace the unobserved values, that is, all the Z_n for $n = r+1, \ldots, N$, with $T + \theta^k$. Therefore, at the E-step we have

$$E(\log f_X(X|\theta)|y, \theta^k) = -N \log \theta - \frac{1}{\theta}\left(\left(\sum_{n=1}^{N} y_n\right) + (N - r)\theta^k\right). \tag{55}$$

The M-step is to maximize this function of θ, which leads to

$$\theta^{k+1} = \left(\left(\sum_{n=1}^{N} y_n\right) + (N - r)\theta^k\right)/N. \tag{56}$$

Let θ^* be a fixed point of this iteration. Then we have

$$\theta^* = \left(\left(\sum_{n=1}^{N} y_n\right) + (N - r)\theta^*\right)/N,$$

so that

$$\theta^* = \frac{1}{r}\sum_{n=1}^{N} y_n,$$

which, as we have seen, is the likelihood maximizer.

A More General Approach

Let X take values in $\mathbb{R}^N$ and $Y = h(X)$ take values in $\mathbb{R}^M$, where $M < N$ and $h : \mathbb{R}^N \to \mathbb{R}^M$ is a (possibly) many-to-one function. Suppose that there is a second function $k : \mathbb{R}^N \to \mathbb{R}^{N-M}$ such that the function

$$G(x) = (h(x), k(x)) = (y, w) = u \tag{57}$$

has inverse $H(y, w) = x$. Denote by $J(y, w)$ the determinant of the Jacobian matrix associated with the transformation G. Let

$$\mathcal{W}(y) = \{w | w = k(x), \text{ and } y = h(x)\}.$$

Then

$$f_Y(y|\theta) = \int_{w \in \mathcal{W}(y)} f_X(H(y, w)) J(y, w) dw. \tag{58}$$

Then we apply the missing-data model for the EM algorithm, with $W = k(X)$ as the missing data.

8 A Multinomial Example

In many applications, the entries of the vector y are independent realizations of a single real-valued or vector-valued random variable V, as they are, at least initially, for finite mixture problems to be considered later. This is not always the case, however, as the following example shows.

A well-known example that was used in [11] and again in [1] to illustrate the EM algorithm concerns a multinomial model taken from genetics. Here there are four cells, with cell probabilities $\frac{1}{2} + \frac{1}{4}\theta_0$, $\frac{1}{4}(1 - \theta_0)$, $\frac{1}{4}(1 - \theta_0)$, and $\frac{1}{4}\theta_0$, for some $\theta_0 \in \Theta = [0, 1]$ to be estimated. The entries of y are the frequencies from a sample size of 197. We then have

$$f_Y(y|\theta) = \frac{197!}{y_1! y_2! y_3! y_4!} (\frac{1}{2} + \frac{1}{4}\theta)^{y_1} (\frac{1}{4}(1 - \theta))^{y_2} (\frac{1}{4}(1 - \theta))^{y_3} (\frac{1}{4}\theta)^{y_4}. \tag{59}$$

It is then supposed that the first of the original four cells can be split into two subcells, with probabilities $\frac{1}{2}$ and $\frac{1}{4}\theta_0$. We then write $y_1 = y_{11} + y_{12}$, and let

$$X = (Y_{11}, Y_{12}, Y_2, Y_3, Y_4), \tag{60}$$

where X has a multinomial distribution with five cells. Note that we do now have $Y = h(X)$.

This example is a popular one in the literature on the EM algorithm (see [11] for citations). It is never suggested that the splitting of the first group into two subgroups is motivated by the demands of the genetics theory itself. As stated in [1], the motivation for the splitting is to allow us to view the two random variables $Y_{12} + Y_4$ and $Y_2 + Y_3$ as governed by a binomial distribution; that is, we can view the value of $y_{12} + y_4$ as the number of heads and the value $y_2 + y_3$ as the number of tails that occur in the flipping of a biased coin $y_{12} + y_4 + y_2 + y_3$ times. This simplifies the calculation of the likelihood maximizer.

9 The Example of Finite Mixtures

We say that a random vector V taking values in $\mathbb{R}^D$ is a *finite mixture* if, for $j = 1, \ldots, J$, f_j is a probability density function or probability function, $\theta_j \geq 0$ is a weight, the θ_j sum to one, and the probability density function or probability function for V is

$$f_V(v|\theta) = \sum_{j=1}^{J} \theta_j f_j(v). \tag{61}$$

The value of D is unimportant, and for simplicity, we shall assume that $D = 1$.

We draw N independent samples of V, denoted v_n, and let y_n, the nth entry of the vector y, be the number v_n. To create the preferred data we assume that for each n, the number v_n is a sample of the random variable V_n whose pdf or pf is f_{j_n}, where the probability that $j_n = j$ is θ_j. We then let the N entries of the preferred data X be the indices j_n. The conditional distribution of Y, given X, is clearly independent of the parameter vector θ and is given by

$$f_{Y|X}(y|x, \theta) = \prod_{n=1}^{N} f_{j_n}(y_n);$$

therefore, X is acceptable. Note that we cannot recapture the entries of y from those of x, so the model $Y = h(X)$ does not hold here. Note also that although the vector y is taken originally to be a vector whose entries are independently drawn samples from V, when we create the preferred data X, we change our view of y. Now each entry of y is governed by a different distribution, so y is no longer viewed as a vector of independent sample values of a single random vector.

10 The EM and the Kullback-Leibler Distance

We illustrate the usefulness of acceptability and reformulate the M-step in terms of cross-entropy or Kullback-Leibler distance minimization.

Using Acceptable Data

The assumption that the data X is acceptable helps simplify the theoretical discussion of the EM algorithm.

For any preferred X the M-step of the EM algorithm, in the continuous case, is to maximize the function

$$\int f_{X|Y}(x|y, \theta^k) \log f_X(x|\theta) dx, \tag{62}$$

over $\theta \in \Theta$; the integral is replaced by a sum in the discrete case. For notational convenience we let

$$b(\theta^k) = f_{X|Y}(x|y, \theta^k), \tag{63}$$

and

$$f(\theta) = f_X(x|\theta); \tag{64}$$

both functions are functions of the vector variable x. Then the M-step is equivalent to minimizing the Kullback-Leibler or cross-entropy distance

$$\begin{aligned} KL(b(\theta^k), f(\theta)) &= \int f_{X|Y}(x|y, \theta^k) \log\Big(\frac{f_{X|Y}(x|y, \theta^k)}{f_X(x|\theta)}\Big) dx \\ &= \int f_{X|Y}(x|y, \theta^k) \log\Big(\frac{f_{X|Y}(x|y, \theta^k)}{f_X(x|\theta)}\Big) \\ &\quad + f_X(x|\theta) - f_{X|Y}(x|y, \theta^k) dx. \end{aligned} \tag{65}$$

This holds since both $f_X(x|\theta)$ and $f_{X|Y}(x|y, \theta^k)$ are probability density functions or probabilities.

For acceptable X we have

$$\begin{aligned} \log f_{X,Y}(x, y|\theta) &= \log f_{X|Y}(x|y, \theta) + \log f_Y(y|\theta) \\ &= \log f_{Y|X}(y|x) + \log f_X(x|\theta). \end{aligned} \tag{66}$$

Therefore,

$$\begin{aligned} \log f_Y(y|\theta^{k+1}) - \log f_Y(y|\theta) &= KL(b(\theta^k), f(\theta)) - KL(b(\theta^k), f(\theta^{k+1})) \\ &\quad + KL(b(\theta^k), b(\theta^{k+1})) - KL(b(\theta^k), b(\theta)). \end{aligned} \tag{67}$$

Since $\theta = \theta^{k+1}$ minimizes $KL(b(\theta^k), f(\theta))$, we have that

$$\begin{aligned} \log f_Y(y|\theta^{k+1}) - \log f_Y(y|\theta^k) &= KL(b(\theta^k), f(\theta^k)) \\ &\quad - KL(b(\theta^k), f(\theta^{k+1})) \\ &\quad + KL(b(\theta^k), b(\theta^{k+1})) \geq 0. \end{aligned} \tag{68}$$

This tells us, once again, that the sequence of likelihood values $\{\log f_Y(y|\theta^k)\}$ is increasing and that the sequence of its negatives, $\{-\log f_Y(y|\theta^k)\}$, is decreasing. Since we assume that there is a maximizer θ_{ML} of the likelihood, the sequence $\{-\log f_Y(y|\theta^k)\}$ is also bounded below and the sequences $\{KL(b(\theta^k), b(\theta^{k+1}))\}$ and $\{KL(b(\theta^k), f(\theta^k)) - KL(b(\theta^k), f(\theta^{k+1}))\}$ converge to zero.

Without some notion of convergence in the parameter space Θ, we cannot conclude that $\{\theta^k\}$ converges to a maximum likelihood estimate θ_{ML}. Without some additional assumptions, we cannot even conclude that the functions $f(\theta^k)$ converge to $f(\theta_{ML})$.

11 The Approach of Csiszár and Tusnády

For acceptable X the M-step of the EM algorithm is to minimize the function $KL(b(\theta^k), f(\theta))$ over $\theta \in \Theta$ to get θ^{k+1}. To put the EM algorithm into the framework of the *alternating-minimization* approach of Csiszár and Tusnády [12], we need to view the M-step in a slightly different way; the problem is that, for the continuous case, having found θ^{k+1}, we do not then minimize $KL(b(\theta), f(\theta^{k+1}))$ at the next step.

The Framework of Csiszár and Tusnády

Following [12], we take $\Psi(p, q)$ to be a real-valued function of the variables $p \in P$ and $q \in Q$, where P and Q are arbitrary sets. Minimizing $\Psi(p, q^n)$ gives p^n and minimizing $\Psi(p^n, q)$ gives q^{n+1}, so that

$$\Psi(p^n, q^n) \geq \Psi(p^n, q^{n+1}) \geq \Psi(p^{n+1}, q^{n+1}). \tag{69}$$

The objective is to find $(\hat{p}, \hat{q})$ such that

$$\Psi(p, q) \geq \Psi(\hat{p}, \hat{q}),$$

for all p and q. In order to show that $\{\Psi(p^n, q^n)\}$ converges to

$$d = \inf_{p \in P, q \in Q} \Psi(p, q)$$

the authors of [12] assume the three- and four-point properties.

If there is a nonnegative function $\Delta : P \times P \to \mathbb{R}$ such that

$$\Psi(p, q^{n+1}) - \Psi(p^{n+1}, q^{n+1}) \geq \Delta(p, p^{n+1}), \tag{70}$$

then the *three-point property* holds. If

$$\Delta(p, p^n) + \Psi(p, q) \geq \Psi(p, q^{n+1}), \tag{71}$$

for all p and q, then the *four-point property* holds. Combining these two inequalities, we have

$$\Delta(p, p^n) - \Delta(p, p^{n+1}) \geq \Psi(p^{n+1}, q^{n+1}) - \Psi(p, q). \tag{72}$$

From the inequality in (72) it follows easily that the sequence $\{\Psi(p^n, q^n)\}$ converges to d. Suppose this is not the case. Then there are p', q', and $D > d$ with

$$\Psi(p^n, q^n) \geq D > \Psi(p', q') \geq d.$$

From Eq. (72) we have

$$\Delta(p', p^n) - \Delta(p', p^{n+1}) \geq \Psi(p^{n+1}, q^{n+1}) - \Psi(p', q') \geq D - \Psi(p', q') > 0.$$

But since $\{\Delta(p', p^n)\}$ is a decreasing sequence of positive quantities, successive differences must converge to zero; that is, $\{\Psi(p^{n+1}, q^{n+1})\}$ must converge to $\Psi(p', q')$, which is a contradiction.

The *five-point property* of [12] is obtained by combining (70) and (71):

$$\Psi(p, q) + \Psi(p, q^{n-1}) \geq \Psi(p, q^n) + \Psi(p^n, q^{n-1}). \tag{73}$$

Note that the five-point property does not involve the second function $\Delta(p', p)$. However, assuming that the five-point property holds, it is possible to define $\Delta(p', p)$ so that both the three- and four-point properties hold. Assuming the five-point property, we have

$$\Psi(p, q^{n-1}) - \Psi(p, q^n) \geq \Psi(p^n, q^n) - \Psi(p, q), \tag{74}$$

from which we can show easily that $\{\Psi(p^n, q^n)\}$ converges to d.

Alternating Minimization for the EM Algorithm

Assume that X is acceptable. We define the function $F(\theta)$ to be

$$F(\theta) = \int f_{X|Y}(x|y, \theta) \log f_{Y|X}(y|x) dx, \tag{75}$$

for the continuous case, with a sum replacing the integral for the discrete case. Using the identities

$$\begin{aligned} f_{X,Y}(x, y|\theta) &= f_{X|Y}(x|y, \theta) f_Y(y|\theta) \\ &= f_{Y|X}(y|x, \theta) f_X(x|\theta) = f_{Y|X}(y|x) f_X(x|\theta), \end{aligned}$$

we then have

$$\log f_Y(y|\theta) = F(\theta') + KL(b(\theta'), b(\theta)) - KL(b(\theta'), f(\theta)), \tag{76}$$

for any parameter values θ and θ'. With the choice of $\theta' = \theta$, we have

$$\log f_Y(y|\theta) = F(\theta) - KL(b(\theta), f(\theta)). \tag{77}$$

Therefore, subtracting Eq. 77 from Eq. 76, we get

$$\begin{aligned}\Big(KL(b(\theta'), f(\theta)) - F(\theta')\Big) - \Big(KL(b(\theta), f(\theta)) - F(\theta)\Big) \\ = KL(b(\theta'), b(\theta)).\end{aligned} \tag{78}$$

Now we can put the EM algorithm into the alternating-minimization framework.

Define

$$\Psi(b(\theta'), f(\theta)) = KL(b(\theta'), f(\theta)) - F(\theta'). \tag{79}$$

We know from Eq. (78) that

$$\Psi(b(\theta'), f(\theta)) - \Psi(b(\theta), f(\theta)) = KL(b(\theta'), b(\theta)). \tag{80}$$

Therefore, we can say that the M-step of the EM algorithm is to minimize $\Psi(b(\theta^k), f(\theta))$ over $\theta \in \Theta$ to get θ^{k+1} and that minimizing $\Psi(b(\theta), f(\theta^{k+1}))$ gives us $\theta = \theta^{k+1}$ again. Because the EM algorithm can be viewed as an alternating minimization method, it is also a particular case of the sequential unconstrained minimization techniques [13] and of "optimization transfer," [4].

With the choice of

$$\Delta(b(\theta'), b(\theta)) = KL(b(\theta'), b(\theta)),$$

Eq. (80) becomes

$$\Psi(b(\theta'), f(\theta)) - \Psi(b(\theta), f(\theta)) = \Delta(b(\theta'), b(\theta)), \tag{81}$$

which is the three-point property.

With $P = \mathcal{B}(\Theta)$ and $Q = \mathcal{F}(\Theta)$, the collections of all functions $b(\theta)$ and $f(\theta)$, respectively, we can view the EM algorithm as alternating minimization of the function $\Psi(p, q)$, over $p \in P$ and $q \in Q$. As we have seen, the three-point property holds. What about the four-point property?

The Kullback-Leibler distance is an example of a jointly convex Bregman distance. According to a lemma of Eggermont and LaRiccia [14, 15], the four-point property holds for alternating minimization of such distances, using $\Delta(p', p) = KL(p', p)$, provided that the sets P and Q are closed and convex subsets of $\mathbb{R}^N$. In the continuous case of the EM algorithm, we are not performing alternating minimization on the function $KL(b(\theta), f(\theta'))$, but on $KL(b(\theta), f(\theta')) + F(\theta)$. In the discrete case, whenever $Y = h(X)$, the function $F(\theta)$ is always zero, so we are performing alternating minimization on the KL distance $KL(b(\theta), f(\theta'))$. In [16] the authors consider the problem of minimizing a function of the form

$$\Lambda(p,q) = \phi(p) + \psi(q) + D_g(p,q), \tag{82}$$

where ϕ and ψ are convex and differentiable on $\mathbb{R}^J$, D_g is a Bregman distance, and $P = Q$ is the interior of the domain of g. In [13] it was shown that when D_g is jointly convex, the function $\Lambda(p,q)$ has the five-point property of [12], which is equivalent to the three- and four-point properties taken together. In some particular instances, the collection of the functions $f(\theta)$ is a convex subset of $\mathbb{R}^J$, as well, so the three- and four-point properties hold.

As we saw previously, to have $\Psi(p^n, q^n)$ converging to d, it is sufficient that the five-point property hold. It is conceivable, then, that the five-point property may hold for Bregman distances under somewhat more general conditions than those employed in the Eggermont-LaRiccia Lemma.

The five-point property for the EM case is the following:

$$KL(b(\theta), f(\theta^k)) - KL(b(\theta), f(\theta^{k+1})) \geq \Big(KL(b(\theta^k), f(\theta^k)) - F(\theta^k)\Big) - \Big(KL(b(\theta), f(\theta)) - F(\theta)\Big). \tag{83}$$

12 Sums of Independent Poisson Random Variables

The EM is often used with aggregated data. The case of sums of independent Poisson random variables is particularly important.

Poisson Sums

Let $X_1, \ldots, X_N$ be independent Poisson random variables with expected value $E(X_n) = \lambda_n$. Let X be the random vector with X_n as its entries, λ the vector whose entries are the λ_n, and $\lambda_+ = \sum_{n=1}^N \lambda_n$. Then the probability function for X is

$$f_X(x|\lambda) = \prod_{n=1}^N \lambda_n^{x_n} \exp(-\lambda_n)/x_n! = \exp(-\lambda_+) \prod_{n=1}^N \lambda_n^{x_n}/x_n! \,. \tag{84}$$

Now let $Y = \sum_{n=1}^N X_n$. Then, the probability function for Y is

$$\text{Prob}(Y = y) = \text{Prob}(X_1 + \ldots + X_N = y) \tag{85}$$

$$= \sum_{x_1 + \ldots x_N = y} \exp(-\lambda_+) \prod_{n=1}^N \lambda_n^{x_n}/x_n! \,. \tag{86}$$

As we shall see shortly, we have

$$\sum_{x_1+\ldots x_N=y} \exp(-\lambda_+) \prod_{n=1}^{N} \lambda_n^{x_n}/x_n! = \exp(-\lambda_+)\lambda_+^y/y!. \tag{87}$$

Therefore, Y is a Poisson random variable with $E(Y) = \lambda_+$.

When we observe an instance of Y, we can consider the conditional distribution $f_{X|Y}(x|y,\lambda)$ of $\{X_1,\ldots,X_N\}$, subject to $y = X_1 + \ldots + X_N$. We have

$$f_{X|Y}(x|y,\lambda) = \frac{y!}{x_1!\ldots x_N!}\Big(\frac{\lambda_1}{\lambda_+}\Big)^{x_1}\ldots\Big(\frac{\lambda_N}{\lambda_+}\Big)^{x_N}. \tag{88}$$

This is a *multinomial distribution*.

Given y and λ, the conditional expected value of X_n is then

$$E(X_n|y,\lambda) = y\lambda_n/\lambda_+.$$

To see why this is true, consider the marginal conditional distribution $f_{X_1|Y}(x_1|y,\lambda)$ of X_1, conditioned on y and λ, which we obtain by holding x_1 fixed and summing over the remaining variables. We have

$$f_{X_1|Y}(x_1|y,\lambda) = \frac{y!}{x_1!(y-x_1)!}\Big(\frac{\lambda_1}{\lambda_+}\Big)^{x_1}\Big(\frac{\lambda'_+}{\lambda_+}\Big)^{y-x_1} \sum_{x_2+\ldots+x_N=y-x_1} \frac{(y-x_1)!}{x_2!\ldots x_N!}\prod_{n=2}^{N}\Big(\frac{\lambda_n}{\lambda'_+}\Big)^{x_n},$$

where

$$\lambda'_+ = \lambda_+ - \lambda_1.$$

As we shall show shortly,

$$\sum_{x_2+\ldots+x_N=y-x_1} \frac{(y-x_1)!}{x_2!\ldots x_N!}\prod_{n=2}^{N}\Big(\frac{\lambda_n}{\lambda'_+}\Big)^{x_n} = 1,$$

so that

$$f_{X_1|Y}(x_1|y,\lambda) = \frac{y!}{x_1!(y-x_1)!}\Big(\frac{\lambda_1}{\lambda_+}\Big)^{x_1}\Big(\frac{\lambda'_+}{\lambda_+}\Big)^{y-x_1}.$$

The random variable X_1 is equivalent to the random number of heads showing in y flips of a coin, with the probability of heads given by λ_1/λ_+. Consequently, the conditional expected value of X_1 is $y\lambda_1/\lambda_+$, as claimed. In the next subsection we look more closely at the multinomial distribution.

The Multinomial Distribution

When we expand the quantity $(a_1 + \ldots + a_N)^y$, we obtain a sum of terms, each having the form $a_1^{x_1} \ldots a_N^{x_N}$, with $x_1 + \ldots + x_N = y$. How many terms of the same form are there? There are N variables a_n. We are to use x_n of the a_n, for each $n = 1, \ldots, N$, to get $y = x_1 + \ldots + x_N$ factors. Imagine y blank spaces, each to be filled in by a variable as we do the selection. We select x_1 of these blanks and mark them a_1. We can do that in $\binom{y}{x_1}$ ways. We then select x_2 of the remaining blank spaces and enter a_2 in them; we can do this in $\binom{y-x_1}{x_2}$ ways. Continuing in this way, we find that we can select the N factor types in

$$\binom{y}{x_1}\binom{y-x_1}{x_2}\ldots\binom{y-(x_1+\ldots+x_{N-2})}{x_{N-1}} \tag{89}$$

ways or in

$$\frac{y!}{x_1!(y-x_1)!}\cdots\frac{(y-(x_1+\ldots+x_{N-2}))!}{x_{N-1}!(y-(x_1+\ldots+x_{N-1}))!} = \frac{y!}{x_1!\ldots x_N!}. \tag{90}$$

This tells us in how many different sequences the factor variables can be selected. Applying this, we get the multinomial theorem:

$$(a_1 + \ldots + a_N)^y = \sum_{x_1+\ldots+x_N=y} \frac{y!}{x_1!\ldots x_N!} a_1^{x_1}\ldots a_N^{x_N}. \tag{91}$$

Select $a_n = \lambda_n/\lambda_+$. Then,

$$1 = 1^y = \left(\frac{\lambda_1}{\lambda_+} + \ldots + \frac{\lambda_N}{\lambda_+}\right)^y \tag{92}$$

$$= \sum_{x_1+\ldots+x_N=y} \frac{y!}{x_1!\ldots x_N!}\left(\frac{\lambda_1}{\lambda_+}\right)^{x_1}\ldots\left(\frac{\lambda_N}{\lambda_+}\right)^{x_N}. \tag{93}$$

From this we get

$$\sum_{x_1+\ldots x_N=y} \exp(-\lambda_+)\prod_{n=1}^{N}\lambda_n^{x_n}/x_n! = \exp(-\lambda_+)\lambda_+^y/y!\ . \tag{94}$$

13 Poisson Sums in Emission Tomography

Sums of Poisson random variables and the problem of complete versus incomplete data arise in *single-photon computed emission tomography* (SPECT) [17].

The SPECT Reconstruction Problem

In their 1976 paper, Rockmore and Makovski [18] suggested that the problem of reconstructing a tomographic image be viewed as statistical parameter estimation. Shepp and Vardi [19] expanded on this idea and suggested that the EM algorithm discussed by Dempster, Laird, and Rubin [11] should be used for the reconstruction. The region of interest within the body of the patient is discretized into J pixels (or voxels), with $\lambda_j \geq 0$ the unknown amount of radionuclide within the jth pixel; we assume that λ_j is also the expected number of photons emitted from the jth pixel during the scanning time. Emitted photons are detected at any one of I detectors outside the body, with $y_i > 0$ the photon count at the ith detector. The probability that a photon emitted at the jth pixel will be detected at the ith detector is P_{ij}, which we assume is known; the overall probability of detecting a photon emitted from the jth pixel is $s_j = \sum_{i=1}^{I} P_{ij} > 0$.

The Preferred Data

For each i and j, the random variable X_{ij} is the number of photons emitted from the jth pixel and detected at the ith detector; the X_{ij} are assumed to be independent and $P_{ij}\lambda_j$-Poisson. With x_{ij} a realization of X_{ij}, the vector x with components x_{ij} is our preferred data. The pdf for this preferred data is a probability vector, with

$$f_X(x|\lambda) = \prod_{i=1}^{I}\prod_{j=1}^{J} \exp^{-P_{ij}\lambda_j}(P_{ij}\lambda_j)^{x_{ij}}/x_{ij}! \,. \tag{95}$$

Given an estimate λ^k of the vector λ and the restriction that $Y_i = \sum_{j=1}^{J} X_{ij}$, the random variables $X_{i1}, \ldots, X_{iJ}$ have the multinomial distribution

$$\text{Prob}(x_{i1}, \ldots, x_{iJ}) = \frac{y_i!}{x_{i1}! \cdots x_{iJ}!} \prod_{j=1}^{J} \left(\frac{P_{ij}\lambda_j}{(P\lambda)_i}\right)^{x_{ij}}.$$

Therefore, the conditional expected value of X_{ij}, given y and λ^k, is

$$E(X_{ij}|y, \lambda^k) = \lambda_j^k P_{ij}\left(\frac{y_i}{(P\lambda^k)_i}\right),$$

and the conditional expected value of the random variable

$$\log f_X(X|\lambda) = \sum_{i=1}^{I}\sum_{j=1}^{J}(-P_{ij}\lambda_j) + X_{ij}\log(P_{ij}\lambda_j) + \text{constants}$$

becomes

$$E(\log f_X(X|\lambda)|y, \lambda^k) = \sum_{i=1}^{I}\sum_{j=1}^{J}\left((-P_{ij}\lambda_j) + \lambda_j^k P_{ij}\left(\frac{y_i}{(P\lambda^k)_i}\right)\log(P_{ij}\lambda_j)\right),$$

omitting terms that do not involve the parameter vector λ. In the EM algorithm, we obtain the next estimate λ^{k+1} by maximizing $E(\log f_X(X|\lambda)|y, \lambda^k)$.

The log likelihood function for the preferred data X (omitting constants) is

$$LL_x(\lambda) = \sum_{i=1}^{I} \sum_{j=1}^{J} \Big(- P_{ij}\lambda_j + X_{ij} \log(P_{ij}\lambda_j) \Big). \tag{96}$$

Of course, we do not have the complete data.

The Incomplete Data

What we do have are the y_i, values of the random variables

$$Y_i = \sum_{j=1}^{J} X_{ij}; \tag{97}$$

this is the given data. These random variables are also independent and $(P\lambda)_i$-Poisson, where

$$(P\lambda)_i = \sum_{j=1}^{J} P_{ij}\lambda_j.$$

The log likelihood function for the given data is

$$LL_y(\lambda) = \sum_{i=1}^{I} \Big(- (P\lambda)_i + y_i \log((P\lambda)_i) \Big). \tag{98}$$

Maximizing $LL_x(\lambda)$ in Eq. (96) is easy, while maximizing $LL_y(\lambda)$ in Eq. (98) is harder and requires an iterative method.

The EM algorithm involves two steps: in the E-step we compute the conditional expected value of $LL_x(\lambda)$, conditioned on the data vector y and the current estimate λ^k of λ; in the M-step we maximize this conditional expected value to get the next λ^{k+1}. Putting these two steps together, we have the following EMML iteration:

$$\lambda_j^{k+1} = \lambda_j^k s_j^{-1} \sum_{i=1}^{I} P_{ij} \frac{y_i}{(P\lambda^k)_i}. \tag{99}$$

For any positive starting vector λ^0, the sequence $\{\lambda^k\}$ converges to a maximizer of $LL_y(\lambda)$, over all nonnegative vectors λ.

Note that because we are dealing with finite probability vectors in this example, it is a simple matter to conclude that

$$f_Y(y|\lambda) = \sum_{x \in h^{-1}\{y\}} f_X(x|\lambda). \tag{100}$$

Using the KL Distance

In this subsection we assume, for notational convenience, that the system $y = P\lambda$ has been normalized so that $s_j = 1$ for each j. Maximizing $E(\log f_X(X|\lambda)|y, \lambda^k)$ is equivalent to minimizing $KL(r(\lambda^k), q(\lambda))$, where $r(\lambda)$ and $q(\lambda)$ are I by J arrays with entries

$$r(\lambda)_{ij} = \lambda_j P_{ij}\Big(\frac{y_i}{(P\lambda)_i}\Big),$$

and

$$q(\lambda)_{ij} = \lambda_j P_{ij}.$$

In terms of our previous notation, we identify $r(\lambda)$ with $b(\theta)$ and $q(\lambda)$ with $f(\theta)$. The set $\mathcal{F}(\Theta)$ of all $f(\theta)$ is now a convex set and the four-point property of [12] holds. The iterative step of the EMML algorithm is then

$$\lambda_j^{k+1} = \lambda_j^k \sum_{i=1}^{I} P_{i,j} \frac{y_i}{(P\lambda^k)_i}. \tag{101}$$

The sequence $\{\lambda^k\}$ converges to a maximizer λ_{ML} of the likelihood for any positive starting vector.

As we noted previously, before we can discuss the possible convergence of the sequence $\{\lambda^k\}$ of parameter vectors to a maximizer of the likelihood, it is necessary to have a notion of convergence in the parameter space. For the problem in this section, the parameter vectors λ are nonnegative. Proof of convergence of the sequence $\{\lambda^k\}$ depends heavily on the following [20]:

$$KL(y, P\lambda^k) - KL(y, P\lambda^{k+1}) = KL(r(\lambda^k), r(\lambda^{k+1})) + KL(\lambda^{k+1}, \lambda^k); \tag{102}$$

and

$$KL(\lambda_{ML}, \lambda^k) - KL(\lambda_{ML}, \lambda^{k+1}) \geq KL(y, P\lambda^k) - KL(y, P\lambda_{ML}). \tag{103}$$

14 Nonnegative Solutions for Linear Equations

Any likelihood maximizer λ_{ML} is also a nonnegative minimizer of the KL distance $KL(y, P\lambda)$, so the EMML algorithm can be thought of, more generally, as a method for finding a nonnegative solution (or approximate solution) for a system $y = P\lambda$ of linear equations in which $y_i > 0$ and $P_{ij} \geq 0$ for all indices. This will be helpful when we consider mixture problems.

The General Case

Suppose we want a nonnegative solution x for a system $Ax = b$ of real equations; unless b is positive and A has only nonnegative entries, we cannot use the EMML algorithm directly. We may, however, be able to transform $Ax = b$ to $P\lambda = y$.

Suppose that by rescaling the equations in $Ax = b$, we can make $c_j = \sum_{i=1}^{I} A_{ij} > 0$, for each $j = 1, \ldots, J$, and $b_+ = \sum_{i=1}^{I} b_i > 0$. Now replace A_{ij} with $G_{ij} = A_{ij}/c_j$, and x_j with $z_j = c_j x_j$; then $Gz = Ax = b$ and $\sum_{i=1}^{I} G_{ij} = 1$, for all j. We also know now that $b_+ = z_+ > 0$, so z_+ is now known.

Let U and u be the matrix and column vector whose entries are all one, respectively, and let $t > 0$ be large enough so that all the entries of $B = G + tU$ and $(tz_+)u$ are positive. Now

$$Bz = Gz + (tz_+)u = b + (tz_+)u.$$

We then solve $Bz = b + (tz_+)u$ for z. It follows that $Ax = Gz = b$ and $x \geq 0$. Finally, we let $P = B$, $\lambda = z$, and $y = b + (tb_+)u$.

Regularization

It is often the case, as in tomography, that the entries of the vector y are obtained by measurements and are therefore noisy. Finding an exact solution of $y = P\lambda$ or even minimizing $KL(y, P\lambda)$ may not be advisable in such cases. To obtain an approximate solution that is relatively insensitive to the noise in y,' we *regularize*. One way to do that is to minimize not $KL(y, P\lambda)$, but

$$F_\alpha(\lambda) = (1-\alpha)KL(y, P\lambda) + \alpha KL(p, \lambda), \tag{104}$$

where $\alpha \in (0, 1)$ and $p > 0$ is a prior estimate of the desired λ. The iterative step of the regularized EMML algorithm is now

$$\lambda_j^{k+1} = (1-\alpha)\Big(\lambda_j^k s_j^{-1} \sum_{i=1}^{I} P_{ij} \frac{y_i}{(P\lambda^k)_i}\Big) + \alpha p_j. \tag{105}$$

As was shown in [20], the sequence $\{\lambda^k\}$ converges to a minimizer of $F_\alpha(\lambda)$.

Acceleration

When the system $y = P\lambda$ is large, the EMML algorithm can be slow to converge. One method that has been used to accelerate convergence to a solution is the use of block iteration [21–23].

We begin by writing the index set $\{i = 1, 2, \ldots, I\}$ as the (not necessarily disjoint) union of $B_n, n = 1, 2, \ldots, N$. Of particular interest is the *row-action* EMML, obtained by letting each block be a singleton. At each step of the iteration, we employ only those equations whose index is a member of the current block. We then cycle through the blocks.

An obvious way to impose blocks would seem to be to modify the EMML iteration as follows:

$$\lambda_j^{k+1} = \lambda_j^k s_{n,j}^{-1} \sum_{i \in B_n} P_{ij} \frac{y_i}{(P\lambda^k)_i}, \tag{106}$$

where

$$s_{n,j} = \sum_{i \in B_n} P_{ij}.$$

This does not work, though.

Let $H_i = \{z \geq 0 | (Pz)_i = y_i\}$. Note that for a fixed $x > 0$, we cannot calculate in closed form the vector $z \in H_i$ that minimizes $KL(z, \lambda)$. However, the vector $z = z^i$ in H_i that minimizes the weighted KL distance

$$\sum_{j=1}^{J} P_{ij} KL(z_j, \lambda_j^k)$$

is given by

$$z_j^i = \lambda_j^k \frac{y_i}{(P\lambda^k)_i}. \tag{107}$$

The iterative step of the EMML algorithm can then be interpreted as saying that λ^{k+1} is a weighted arithmetic mean of the z^i; that is,

$$\lambda_j^{k+1} = s_j^{-1} \sum_{i=1}^{I} P_{ij} z_j^i. \tag{108}$$

This suggests a different form for a block-iterative version of the EMML.

For $k = 0, 1, \ldots$, and $n = n(k) = k(\text{mod } N) + 1$, let

$$\lambda_j^{k+1} = (1 - m_n^{-1} s_{nj}) \lambda_j^k + m_n^{-1} \lambda_j^k \sum_{i \in B_n} P_{ij} \frac{y_i}{(P\lambda^k)_i}, \tag{109}$$

where $m_n = \max_j s_{nj}$. This is the *rescaled* block-iterative EMML (RBI-EMML) algorithm. The sequence $\{\lambda^k\}$ converges to a nonnegative solution of the system $y = P\lambda$, for any choice of blocks, whenever the system has a nonnegative solution [21].

When each block is a singleton, that is, $B_n = B_i = \{i\}$, for $i = 1, 2, \ldots, I = N$, the RBI-EMML becomes the EMART algorithm, with the iterative step

$$\lambda_j^{k+1} = (1 - m_i^{-1} P_{ij})\lambda_j^k + \lambda_j^k m_i^{-1} P_{ij} \frac{y_i}{(P\lambda^k)_i}, \tag{110}$$

where $m_i = \max_j P_{ij} > 0$. It is interesting to compare the EMART algorithm with the *multiplicative algebraic reconstruction technique* (MART) [24], which has the iterative step

$$\lambda_j^{k+1} = \lambda_j^k \Big(\frac{y_i}{(P\lambda^k)_i}\Big)^{P_{ij}/m_i}, \tag{111}$$

so that

$$\lambda_j^{k+1} = \Big(\lambda_j^k\Big)^{1-P_{ij}/m_i} \Big(\lambda_j^k \frac{y_i}{(P\lambda^k)_i}\Big)^{p_{ij}/m_i}, \tag{112}$$

or

$$\log \lambda_j^{k+1} = (1 - m_i^{-1} P_{ij}) \log \lambda_j^k + m_i^{-1} P_{ij} \log \Big(\lambda_j^k \frac{y_i}{(P\lambda^k)_i}\Big). \tag{113}$$

The difference between the MART and the EMART is then the difference between a geometric mean and an arithmetic mean.

The simultaneous MART (SMART) is analogous to the EMML and uses all the equations at each step [20, 25, 26]. The iterative step for the SMART is

$$\lambda_j^{k+1} = \lambda_j^k \exp\Big(s_j^{-1} \sum_{i=1}^{I} P_{ij} \log \frac{y_i}{(P\lambda^k)_i}\Big). \tag{114}$$

Block-iterative versions of the MART (RBI-SMART) have been considered by [27] and [21]. When $y = P\lambda$ has nonnegative solutions, the RBI-SMART sequence converges to the nonnegative solution of $y = P\lambda$ for which the cross-entropy $KL(\lambda, \lambda^0)$ is minimized. When there are no nonnegative solutions of $y = P\lambda$, the SMART converges to the nonnegative minimizer of $KL(P\lambda, y)$ for which $KL(\lambda, \lambda^0)$ is minimized [28].

Using Prior Bounds on λ

The EMML algorithm finds an approximate nonnegative solution of $y = P\lambda$. In some applications it is helpful to be able to incorporate upper and lower bounds on the λ [29].

The SMART, EMML, MART, and EMART methods are based on the Kullback-Leibler distance between nonnegative vectors. To impose more general constraints

on the entries of λ, we derive algorithms based on shifted KL distances, also called *Fermi-Dirac generalized entropies*.

For a fixed real vector u, the shifted KL distance $KL(x-u, z-u)$ is defined for vectors x and z having $x_j \geq u_j$ and $z_j \geq u_j$. Similarly, the shifted distance $KL(v-x, v-z)$ applies only to those vectors x and z for which $x_j \leq v_j$ and $z_j \leq v_j$. For $u_j \leq v_j$, the combined distance

$$KL(x-u, z-u) + KL(v-x, v-z)$$

is restricted to those x and z whose entries x_j and z_j lie in the interval $[u_j, v_j]$. Our objective is to mimic the derivation of the SMART and EMML methods, replacing KL distances with shifted KL distances, to obtain algorithms that enforce the constraints $u_j \leq \lambda_j \leq v_j$, for each j. The algorithms that result are the ABMART and ABEMML block-iterative methods. These algorithms were originally presented in [30], in which the vectors u and v were called a and b, hence the names of the algorithms. We shall assume that the entries of the matrix P are nonnegative. We shall denote by B_n, $n = 1, \ldots, N$ a partition of the index set $\{i = 1, \ldots, I\}$ into blocks. For $k = 0, 1, \ldots$ let $n = n(k) = k(\text{mod}\, N) + 1$.

The ABMART Algorithm

We assume that $(Pu)_i \leq y_i \leq (Pv)_i$ and seek a solution of $P\lambda = y$ with $u_j \leq \lambda_j \leq v_j$, for each j. The algorithm begins with an initial vector λ^0 satisfying $u_j \leq \lambda_j^0 \leq v_j$, for each j. Having calculated λ^k, we take

$$\lambda_j^{k+1} = \alpha_j^k v_j + (1-\alpha_j^k) u_j, \tag{115}$$

with $n = n(k)$,

$$\alpha_j^k = \frac{c_j^k \prod^n (d_i^k)^{P_{ij}}}{1 + c_j^k \prod^n (d_i^k)^{A_{ij}}}, \tag{116}$$

$$c_j^k = \frac{(\lambda_j^k - u_j)}{(v_j - \lambda_j^k)}, \tag{117}$$

and

$$d_j^k = \frac{(y_i - (Pu)_i)((Pv)_i - (P\lambda^k)_i)}{((Pv)_i - y_i)((P\lambda^k)_i - (Pu)_i)}, \tag{118}$$

where $\prod^n$ denotes the product over those indices i in $B_{n(k)}$. Notice that at each step of the iteration, λ_j^k is a convex combination of the endpoints u_j and v_j, so that λ_j^k always lies in the interval $[u_j, v_j]$.

We have the following theorem concerning the convergence of the ABMART algorithm:

Theorem 1. *If there is a solution of the system $P\lambda = y$ that satisfies the constraints $u_j \leq \lambda_j \leq v_j$ for each j, then, for any N and any choice of the blocks B_n, the ABMART sequence converges to that constrained solution of $P\lambda = y$ for which the Fermi-Dirac generalized entropic distance from λ to λ^0, given by*

$$KL(\lambda - u, \lambda^0 - u) + KL(v - \lambda, v - \lambda^0),$$

is minimized. If there is no constrained solution of $P\lambda = y$, then, for $N = 1$, the ABMART sequence converges to the minimizer of

$$KL(P\lambda - Pu, y - Pu) + KL(Pv - P\lambda, Pv - y)$$

for which

$$KL(\lambda - u, \lambda^0 - u) + KL(v - \lambda, v - \lambda^0)$$

is minimized.

The proof is in [30].

The ABEMML Algorithm

We make the same assumptions as previously. The iterative step of the ABEMML algorithm is

$$\lambda_j^{k+1} = \alpha_j^k v_j + (1 - \alpha_j^k) u_j, \tag{119}$$

where

$$\alpha_j^k = \gamma_j^k / d_j^k, \tag{120}$$

$$\gamma_j^k = (\lambda_j^k - u_j) e_j^k, \tag{121}$$

$$\beta_j^k = (v_j - \lambda_j^k) f_j^k, \tag{122}$$

$$d_j^k = \gamma_j^k + \beta_j^k, \tag{123}$$

$$e_j^k = \left(1 - \sum_{i \in B_n} P_{ij}\right) + \sum_{i \in B_n} P_{ij} \left(\frac{y_i - (Pu)_i}{(P\lambda^k)_i - (Pu)_i}\right), \tag{124}$$

and

$$f_j^k = \left(1 - \sum_{i \in B_n} P_{ij}\right) + \sum_{i \in B_n} P_{ij} \left(\frac{(Pv)_i - y_i}{(Pv)_i - (P\lambda^k)_i}\right). \tag{125}$$

The following theorem concerns the convergence of the ABEMML algorithm:

Theorem 2. *If there is a solution of the system $P\lambda = y$ that satisfies the constraints $u_j \leq \lambda_j \leq v_j$ for each j, then, for any N and any choice of the blocks B_n, the ABEMML sequence converges to such a constrained solution of $P\lambda = y$. If there is no constrained solution of $P\lambda = y$, then, for $N = 1$, the ABEMML sequence converges to a constrained minimizer of*

$$KL(y - Pu, P\lambda - Pu) + KL(Pv - y, Pv - P\lambda).$$

The proof is found in [30]. In contrast to the ABMART theorem, this is all we can say about the limits of the ABEMML sequences.

15 Finite Mixture Problems

Estimating the combining proportions in probabilistic mixture problems shows that there are meaningful examples of our acceptable-data model, and provides important applications of likelihood maximization.

Mixtures

We say that a random vector V taking values in $\mathbb{R}^D$ is a *finite mixture* [31, 32] if there are probability density functions or probabilities f_j and numbers $\theta_j \geq 0$, for $j = 1, \ldots, J$, such that the probability density function or probability function for V has the form

$$f_V(v|\theta) = \sum_{j=1}^{J} \theta_j f_j(v), \tag{126}$$

for some choice of the $\theta_j \geq 0$ with $\sum_{j=1}^{J} \theta_j = 1$. As previously, we shall assume, without loss of generality, that $D = 1$.

The Likelihood Function

The data are N realizations of the random variable V, denoted v_n, for $n = 1, \ldots, N$, and the given data is the vector $y = (v_1, \ldots, v_N)$. The column vector $\theta = (\theta_1, \ldots, \theta_J)^T$ is the generic parameter vector of mixture combining proportions. The likelihood function is

$$L_y(\theta) = \prod_{n=1}^{N} \Big(\theta_1 f_1(v_n) + \ldots + \theta_J f_J(v_n)\Big). \tag{127}$$

Then the log likelihood function is

$$LL_y(\theta) = \sum_{n=1}^{N} \log\Big(\theta_1 f_1(v_n) + \ldots + \theta_J f_J(v_n)\Big).$$

With u the column vector with entries $u_n = 1/N$, and P the matrix with entries $P_{nj} = f_j(v_n)$, we define

$$s_j = \sum_{n=1}^{N} P_{nj} = \sum_{n=1}^{N} f_j(v_n).$$

Maximizing $LL_y(\theta)$ is equivalent to minimizing

$$F(\theta) = KL(u, P\theta) + \sum_{j=1}^{J}(1 - s_j)\theta_j. \tag{128}$$

A Motivating Illustration

To motivate such mixture problems, we imagine that each data value is generated by first selecting one value of j, with probability θ_j, and then selecting a realization of a random variable governed by $f_j(v)$. For example, there could be J bowls of colored marbles, and we randomly select a bowl, and then randomly select a marble within the selected bowl. For each n the number v_n is the numerical code for the color of the nth marble drawn. In this illustration we are using a mixture of probability functions, but we could have used probability density functions.

The Acceptable Data

We approach the mixture problem by creating acceptable data. We imagine that we could have obtained $x_n = j_n$, for $n = 1, \ldots, N$, where the selection of v_n is governed by the function $f_{j_n}(v)$. In the bowls example, j_n is the number of the bowl from which the nth marble is drawn. The acceptable-data random vector is $X = (X_1, \ldots, X_N)$, where the X_n are independent random variables taking values in the set $\{j = 1, \ldots, J\}$. The value j_n is one realization of X_n. Since our objective is to estimate the true θ_j, the values v_n are now irrelevant. Our ML estimate of the true θ_j is simply the proportion of times $j = j_n$. Given a realization x of X, the conditional pdf or pf of Y does not involve the mixing proportions, so X is acceptable. Notice also that it is not possible to calculate the entries of y from those of x; the model $Y = h(X)$ does not hold.

The Mix-EM Algorithm

Using this acceptable data, we derive the EM algorithm, which we call the Mix-EM algorithm.

With N_j denoting the number of times the value j occurs as an entry of x, the likelihood function for X is

$$L_x(\theta) = f_X(x|\theta) = \prod_{j=1}^{J} \theta_j^{N_j}, \tag{129}$$

and the log likelihood is

$$LL_x(\theta) = \log L_x(\theta) = \sum_{j=1}^{J} N_j \log \theta_j. \tag{130}$$

Then

$$E(\log L_x(\theta)|y, \theta^k) = \sum_{j=1}^{J} E(N_j|y, \theta^k) \log \theta_j. \tag{131}$$

To simplify the calculations in the E-step, we rewrite $LL_x(\theta)$ as

$$LL_x(\theta) = \sum_{n=1}^{N} \sum_{j=1}^{J} X_{nj} \log \theta_j, \tag{132}$$

where $X_{nj} = 1$ if $j = j_n$ and zero otherwise. Then we have

$$E(X_{nj}|y, \theta^k) = \text{prob}\,(X_{nj} = 1|y, \theta^k) = \frac{\theta_j^k f_j(v_n)}{f(v_n|\theta^k)}. \tag{133}$$

The function $E(LL_x(\theta)|y, \theta^k)$ becomes

$$E(LL_x(\theta)|y, \theta^k) = \sum_{n=1}^{N} \sum_{j=1}^{J} \frac{\theta_j^k f_j(v_n)}{f(v_n|\theta^k)} \log \theta_j. \tag{134}$$

Maximizing with respect to θ, we get the iterative step of the Mix-EM algorithm:

$$\theta_j^{k+1} = \frac{1}{N} \theta_j^k \sum_{n=1}^{N} \frac{f_j(v_n)}{f(v_n|\theta^k)}. \tag{135}$$

We know from our previous discussions that since the preferred data X is acceptable, likelihood is non-decreasing for this algorithm. We shall go further now and show that the sequence of probability vectors $\{\theta^k\}$ converges to a maximizer of the likelihood.

Convergence of the Mix-EM Algorithm

As we noted earlier, maximizing the likelihood in the mixture case is equivalent to minimizing

$$F(\theta) = KL(u, P\theta) + \sum_{j=1}^{J}(1 - s_j)\theta_j,$$

over probability vectors θ. It is easily shown that if $\hat{\theta}$ minimizes $F(\theta)$ over all nonnegative vectors θ, then $\hat{\theta}$ is a probability vector. Therefore, we can obtain the maximum likelihood estimate of θ by minimizing $F(\theta)$ over nonnegative vectors θ.

The following theorem is found in [33].

Theorem 3. *Let u be any positive vector, P any nonnegative matrix with $s_j > 0$ for each j, and*

$$F(\theta) = KL(u, P\theta) + \sum_{j=1}^{J}\beta_j KL(\gamma_j, \theta_j).$$

If $s_j + \beta_j > 0$, $\alpha_j = s_j/(s_j + \beta_j)$, and $\beta_j\gamma_j \geq 0$, for all j, then the iterative sequence given by

$$\theta_j^{k+1} = \alpha_j s_j^{-1}\theta_j^k\Big(\sum_{n=1}^{N} P_{n,j}\frac{u_n}{(P\theta^k)_n}\Big) + (1 - \alpha_j)\gamma_j \tag{136}$$

converges to a nonnegative minimizer of $F(\theta)$.

With the choices $u_n = 1/N$, $\gamma_j = 0$, and $\beta_j = 1 - s_j$, the iteration in Eq. (136) becomes that of the Mix-EM algorithm. Therefore, the sequence $\{\theta^k\}$ converges to the maximum likelihood estimate of the mixing proportions.

16 More on Convergence

There is a mistake in the proof of convergence given in [11]. Wu [34] and Boyles [35] attempted to repair the error but also gave examples in which the EM algorithm failed to converge to a global maximizer of likelihood. In Chap. 3 of the book

by McLachlan and Krishnan [1], we find the basic theory of the EM algorithm, including available results on convergence and the rate of convergence. Because many authors rely on Eq. (14), it is not clear that these results are valid in the generality in which they are presented. There appears to be no single convergence theorem that is relied on universally; each application seems to require its own proof of convergence. When the use of the EM algorithm was suggested for SPECT and PET, it was necessary to prove convergence of the resulting iterative algorithm in Eq. (99), as was eventually achieved in a sequence of papers [19, 36–38], and [20]. When the EM algorithm was applied to list-mode data in SPECT and PET [39–41], the resulting algorithm differed slightly from that in Eq. (99) and a proof of convergence was provided in [33]. The convergence theorem in [33] also establishes the convergence of the iteration in Eq. (135) to the maximum-likelihood estimate of the mixing proportions.

17 Open Questions

As we have seen, the conventional formulation of the EM algorithm presents difficulties when probability density functions are involved. We have shown here that the use of acceptable preferred data can be helpful in resolving this issue, but other ways may also be useful.

Proving convergence of the sequence $\{\theta^k\}$ appears to involve the selection of an appropriate topology for the parameter space Θ. While it is common to assume that Θ is a subset of Euclidean space and that the usual norm should be used to define distance, it may be helpful to tailor the metric to the nature of the parameters. In the case of Poisson sums, for example, the parameters are nonnegative vectors, and we found that the cross-entropy distance is more appropriate. Even so, additional assumptions appear necessary before convergence of the $\{\theta^k\}$ can be established. To simplify the analysis, it is often assumed that cluster points of the sequence lie in the interior of the set Θ, which is not a realistic assumption in some applications.

It may be wise to consider, instead, convergence of the functions $f_X(x|\theta^k)$ or maybe even to identify the parameters θ with the functions $f_X(x|\theta)$. Proving convergence to $L_y(\theta_{ML})$ of the likelihood values $L_y(\theta^k)$ is also an option.

18 Conclusion

Difficulties with the conventional formulation of the EM algorithm in the continuous case of probability density functions (pdf) have prompted us to adopt a new definition, that of acceptable data. As we have shown, this model can be helpful in generating EM algorithms in a variety of situations. For the discrete case of probability functions (pf), the conventional approach remains satisfactory. In both cases, the two steps of the EM algorithm can be viewed as alternating minimization of the Kullback-Leibler distance between two sets of parameterized pf or pdf, along

the lines investigated by Csiszár and Tusnády [12]. In order to use the full power of their theory, however, we need the sets to be closed and convex. This does occur in the important special case of sums of independent Poisson random variables, but is not generally the case.

Acknowledgments I wish to thank Professor Paul Eggermont of the University of Delaware for helpful discussions on these matters.

Cross-References

▸ EM Algorithms
▸ Iterative Solution Methods
▸ Large-Scale Inverse Problems in Imaging
▸ Linear Inverse Problems
▸ Mathematical Methods in PET and Spect Imaging

References

1. McLachlan, G., Krishnan, T.: The EM Algorithm and Extensions. Wiley, New York (1997)
2. Meng, X., Pedlow, S.: EM: a bibliographic review with missing articles. In: Proceedings of the Statistical Computing Section, American Statistical Association. American Statistical Association, Alexandria (1992)
3. Meng, X., van Dyk, D.: The EM algorithm- an old folk-song sung to a fast new tune. J. R. Stat. Soc. B **59**(3), 511–567 (1997)
4. Becker, M., Yang, I., Lange, K.: EM algorithms without missing data. Stat. Methods Med. Res. **6**, 38–54 (1997)
5. Byrne, C.: Iterative Optimization in Inverse Problems. Taylor and Francis, Boca Raton (2014)
6. Byrne, C.: Non-stochastic EM algorithms in optimization. J. Nonlinear Convex Anal. (to appear, 2015)
7. Kullback, S., Leibler, R.: On information and sufficiency. Ann. Math. Stat. **22**, 79–86 (1951)
8. Hogg, R., McKean, J., Craig, A.: Introduction to Mathematical Statistics, 6th edn. Prentice Hall, Englewood Cliffs (2004)
9. Byrne, C., Eggermont, P.: EM algorithms. In: Scherzer, O. (ed.) Handbook of Mathematical Methods in Imaging. Springer, New York (2010)
10. Fessler, J., Ficaro, E., Clinthorne, N., Lange, K.: Grouped-coordinate ascent algorithms for penalized-likelihood transmission image reconstruction. IEEE Trans. Med. Imaging **16**(2), 166–175 (1997)
11. Dempster, A., Laird, N., Rubin, D.: Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. B **37**, 1–38 (1977)
12. Csiszár, I., Tusnády, G.: Information geometry and alternating minimization procedures. Stat. Decis. (Suppl. 1), R. Oldenbourg Verlag, München, 205–237 (1984)
13. Byrne, C.: Alternating and sequential unconstrained minimization algorithms. J. Optim. Theory Appl. Electron. **154**(3), doi:10.1007/s1090134-2; hardcopy **156**(2) (2012)
14. Eggermont, P., LaRiccia, V.: Smoothed maximum likelihood density estimation for inverse problems. Ann. Stat. **23**, 199–220 (1995)
15. Eggermont, P., LaRiccia, V.: Maximum Penalized Likelihood Estimation, Volume I: Density Estimation. Springer, New York (2001)

16. Bauschke, H., Combettes, P., Noll, D.: Joint minimization with alternating Bregman proximity operators. Pac. J. Opt. **2**, 401–424 (2006)
17. Wernick, M., Aarsvold, J. (eds.): Emission Tomography: The Fundamentals of PET and SPECT. Elsevier Academic, San Diego (2004)
18. Rockmore, A., Macovski, A.: A maximum likelihood approach to emission image reconstruction from projections. IEEE Trans. Nucl. Sc. NS-**23**, 1428–1432 (1976)
19. Shepp, L., Vardi, Y.: Maximum likelihood reconstruction for emission tomography. IEEE Trans. Med. Imaging MI-**1**, 113–122 (1982)
20. Byrne, C.: Iterative image reconstruction algorithms based on cross-entropy minimization. IEEE Trans. Image Proc. IP-**2**, 96–103 (1993)
21. Byrne, C.: Block-iterative methods for image reconstruction from projections. IEEE Trans. Image Proc. IP-**5**, 792–794 (1996)
22. Byrne, C.: Convergent block-iterative algorithms for image reconstruction from inconsistent data. IEEE Trans. Image Proc. IP-**6**, 1296–1304 (1997)
23. Byrne, C.: Accelerating the EMML algorithm and related iterative algorithms by rescaled block-iterative (RBI) methods. IEEE Trans. Image Proc. IP-**7**, 100–109 (1998)
24. Gordon, R., Bender, R., Herman, G.T.: Algebraic reconstruction techniques (ART) for three-dimensional electron microscopy and x-ray photography. J. Theor. Biol. **29**, 471–481 (1970)
25. Darroch, J., Ratcliff, D.: Generalized iterative scaling for log-linear models. Ann. Math. Stat. **43**, 1470–1480 (1972)
26. Schmidlin, P.: Iterative separation of sections in tomographic scintigrams. Nucl. Med. **15**(1), 1–16 (1972)
27. Censor, Y., Segman, J.: On block-iterative maximization. J. Inf. Opt. Sci. **8**, 275–291 (1987)
28. Byrne, C.: Iterative reconstruction algorithms based on cross-entropy minimization. In: Shepp, L., Levinson, S.E. (eds.) Image Models (and their Speech Model Cousins). IMA Volumes in Mathematics and its Applications vol. 80, pp. 1–11. Springer, New York (1996)
29. Narayanan, M., Byrne, C., King, M.: An interior point iterative maximum-likelihood reconstruction algorithm incorporating upper and lower bounds with application to SPECT transmission imaging. IEEE Trans. Med. Imaging TMI-**20**(4), 342–353 (2001)
30. Byrne, C.: Iterative algorithms for deblurring and deconvolution with constraints. Inverse Probl. **14**, 1455–1467 (1998)
31. Everitt, B., Hand, D.: Finite Mixture Distributions. Chapman and Hall, London (1981)
32. Redner, R., Walker, H.: Mixture densities, maximum likelihood and the EM algorithm. SIAM Rev. **26**(2), 195–239 (1984)
33. Byrne, C.: Likelihood maximization for list-mode emission tomographic image reconstruction. IEEE Trans. Med. Imaging **20**(10), 1084–1092 (2001)
34. Wu, C.F.J.: On the convergence properties of the EM algorithm. Ann. Stat. **11**, 95–103 (1983)
35. Boyles, R.: On the convergence of the EM algorithm. J. R. Stat. Soc. B **45**, 47–50 (1983)
36. Lange, K., Carson, R.: EM reconstruction algorithms for emission and transmission tomography. J. Comput. Assist. Tomogr. **8**, 306–316 (1984)
37. Vardi, Y., Shepp, L.A., Kaufman, L.: A statistical model for positron emission tomography. J. Am. Stat. Assoc. **80**, 8–20 (1985)
38. Lange, K., Bahn, M., Little, R.: A theoretical study of some maximum likelihood algorithms for emission and transmission tomography. IEEE Trans. Med. Imaging MI-**6**(2), 106–114 (1987)
39. Barrett, H., White, T., Parra, L.: List-mode likelihood. J. Opt. Soc. Am. A **14**, 2914–2923 (1997)
40. Parra, L., Barrett, H.: List-mode likelihood: EM algorithm and image quality estimation demonstrated on 2-D PET. IEEE Trans. Med. Imaging **17**, 228–235 (1998)
41. Huesman, R., Klein, G., Moses, W., Qi, J., Ruetter, B., Virador, P.: List-mode maximum likelihood reconstruction applied to positron emission mammography (PEM) with irregular sampling. IEEE Trans. Med. Imaging **19**(5), 532–537 (2000)

Iterative Solution Methods

Martin Burger , Barbara Kaltenbacher, and Andreas Neubauer

Contents

1 Introduction 432
2 Preliminaries 432
Conditions on F 432
Source Conditions 434
Stopping Rules 434
3 Gradient Methods 435
Nonlinear Landweber Iteration 435
Landweber Iteration in Hilbert Scales 442
Steepest Descent and Minimal Error Method 446
Further Literature on Gradient Methods 446
4 Newton Type Methods 447
Levenberg–Marquardt and Inexact Newton Methods 447
Further Literature on Inexact Newton Methods 451
Iteratively Regularized Gauss–Newton Method 451
Further Literature on Gauss–Newton Type Methods 455
5 Nonstandard Iterative Methods 458
Kaczmarz and Splitting Methods 458
EM Algorithms 460
Bregman Iterations 464
6 Conclusion 465
Cross-References 465
References 466

M. Burger (✉)
Institute for Computational and Applied Mathematics, University of Münster, Münster, Germany
e-mail: martin.burger@uni-muenster.de

B. Kaltenbacher
Institut für Mathematik, Alpen-Adria-Universität Klagenfurt, Klagenfurt, Austria
e-mail: barbara.kaltenbacher@aau.at

A. Neubauer
Industrial Mathematics Institute, Johannes Kepler University Linz, Linz, Austria
e-mail: neubauer@indmath.uni-linz.ac.at

© Springer Science+Business Media New York 2015

O. Scherzer (ed.), *Handbook of Mathematical Methods in Imaging*,
DOI 10.1007/978-1-4939-0790-8_9

Abstract

This chapter deals with iterative methods for nonlinear ill-posed problems. We present gradient and Newton type methods as well as nonstandard iterative algorithms such as Kaczmarz, expectation maximization, and Bregman iterations.

Our intention here is to cite convergence results in the sense of regularization and to provide further references to the literature.

1 Introduction

This chapter will be devoted to the iterative solution of inverse problems formulated as nonlinear operator equations

$$F(x) = y\,, \tag{1}$$

where $F : \mathcal{D}(F) \to \mathcal{Y}$ with domain $\mathcal{D}(F) \subseteq \mathcal{X}$. The exposition will be mainly restricted to the case of $\mathcal{X}$ and $\mathcal{Y}$ being Hilbert spaces with inner products $\langle \cdot\,, \cdot \rangle$ and norms $\|\cdot\|$. Some references for the Banach space case will be given.

We will assume attainability of the exact data y in a ball $\mathcal{B}_\rho(x_0)$, i.e., the equation $F(x) = y$ is solvable in $\mathcal{B}_\rho(x_0)$. The element x_0 is an initial guess which may incorporate a-priori knowledge of an exact solution.

The actually available data y^δ will in practice usually be contaminated with noise for which we here use a deterministic model, i.e.,

$$\|y^\delta - y\| \leq \delta\,, \tag{2}$$

where the noise level δ is assumed to be known. For a convergence analysis with stochastic noise, see the references in section "Further Literature on Gauss–Newton Type Methods".

2 Preliminaries

Conditions on F

For the proofs of well-definedness and local convergence of the iterative methods considered here we need several conditions on the operator F. Basically, we inductively show that the iterates remain in a neighborhood of the initial guess. Hence, to guarantee applicability of the forward operator to these iterates, we assume that

$$\mathcal{B}_{2\rho}(x_0) \subseteq \mathcal{D}(F) \tag{3}$$

for some $\rho > 0$.

Moreover, we need that F is continuously Fréchet-differentiable, that $\|F'(x)\|$ is uniformly bounded with respect to $x \in \mathcal{B}_{2\rho}(x_0)$, and that problem (1) is properly scaled, i.e., certain parameters occurring in the iterative methods have to be chosen appropriately in dependence of this uniform bound.

The assumption that F' is Lipschitz continuous,

$$\|F'(\tilde{x}) - F'(x)\| \leq L\,\|\tilde{x} - x\|\,, \qquad x, \tilde{x} \in \mathcal{B}_{2\rho}(x_0)\,, \tag{4}$$

that is often used to show convergence of iterative methods for well-posed problems, implies that

$$\|F(\tilde{x}) - F(x) - F'(x)(\tilde{x} - x)\| \leq c\,\|\tilde{x} - x\|^2\,, \qquad x, \tilde{x} \in \mathcal{B}_{2\rho}(x_0)\,. \tag{5}$$

However, this Taylor remainder estimate is too weak for the ill-posed situation unless the solution is sufficiently smooth (see, e.g., case (ii) in Theorem 9 below). An assumption on F that can often be found in the literature on nonlinear ill-posed problems is the *tangential cone condition*

$$\begin{aligned} &\|F(x) - F(\tilde{x}) - F'(x)(x - \tilde{x})\| \leq \eta \\ &\qquad \|F(x) - F(\tilde{x})\|\,, \quad \eta < \tfrac{1}{2}\,, \quad x, \tilde{x} \in \mathcal{B}_{2\rho}(x_0) \subseteq \mathcal{D}(F)\,, \end{aligned} \tag{6}$$

which implies that

$$\frac{1}{1+\eta}\,\|F'(x)(\tilde{x} - x)\| \leq \|F(\tilde{x}) - F(x)\| \leq \frac{1}{1-\eta}\,\|F'(x)(\tilde{x} - x)\|$$

for all $x, \tilde{x} \in \mathcal{B}_{2\rho}(x_0)$. One can even prove (see [70, Proposition 2.1]).

Proposition 1. *Let $\rho, \varepsilon > 0$ be such that*

$$\begin{aligned} &\|F(x) - F(\tilde{x}) - F'(x)(x - \tilde{x})\| \leq c(x, \tilde{x}) \\ &\qquad \|F(x) - F(\tilde{x})\|\,, \quad x, \tilde{x} \in \mathcal{B}_{\rho}(x_0) \subseteq \mathcal{D}(F), \end{aligned}$$

for some $c(x, \tilde{x}) \geq 0$, where $c(x, \tilde{x}) < 1$ if $\|x - \tilde{x}\| \leq \varepsilon$.

(i) Then for all $x \in \mathcal{B}_{\rho}(x_0)$

$$M_x := \{\tilde{x} \in \mathcal{B}_{\rho}(x_0) \,:\, F(\tilde{x}) = F(x)\} = x + \mathcal{N}(F'(x)) \cap \mathcal{B}_{\rho}(x_0)$$

and $\mathcal{N}(F'(x)) = \mathcal{N}(F'(\tilde{x}))$ for all $\tilde{x} \in M_x$. Moreover,

$$\mathcal{N}(F'(x)) \supseteq \{t(\tilde{x} - x) \,:\, \tilde{x} \in M_x, t \in \mathbb{R}\}\,,$$

where instead of $\supseteq$ equality holds if $x \in \overset{\circ}{\mathcal{B}}_{\rho}(x_0)$.

(ii) If $F(x) = y$ is solvable in $\mathcal{B}_\rho(x_0)$, then a unique x_0-minimum-norm solution exists. It is characterized as the solution $x^\dagger$ of $F(x) = y$ in $\mathcal{B}_\rho(x_0)$ satisfying the condition

$$x^\dagger - x_0 \in \mathcal{N}(F'(x^\dagger))^\perp \subseteq \mathcal{X}\,. \tag{7}$$

If $F(x) = y$ is solvable in $\mathcal{B}_\rho(x_0)$ but a condition like (6) is not satisfied, then at least existence (but no uniqueness) of an x_0-minimum-norm solution is guaranteed provided that F is weakly sequentially closed (see [36, Chapter 10]).

For the proofs of convergence rates one even needs stronger conditions on F' than condition (6).

Source Conditions

It is well known by now that the convergence of regularized solutions can be arbitrarily slow. Rates can only be proven if the exact solution $x^\dagger$ satisfies some regularity assumptions, so-called source conditions. They are usually of Hölder-type, i.e.,

$$x^\dagger - x_0 = (F'(x^\dagger)^* F'(x^\dagger))^\mu v\,, \quad v \in \mathcal{N}(F'(x^\dagger))^\perp \tag{8}$$

for some exponent $\mu > 0$. Due to typical smoothing properties of the linearized forward operator $F'(x^\dagger)$, they can be interpreted as smoothness assumptions on the initial error $x^\dagger - x_0$.

Since (8) is usually too strong for severely ill-posed problems, logarithmic source conditions, i.e.,

$$\begin{aligned} x^\dagger - x_0 &= f_\mu^L(F'(x^\dagger)^* F'(x^\dagger))v, \quad \mu > 0, \quad v \in \mathcal{N}(F'(x^\dagger))^\perp, \\ f_\mu^L(\lambda) &:= (-\ln(\lambda c_L^{-1}))^{-\mu}, \quad c_L > c_s^2\,, \end{aligned} \tag{9}$$

have been considered by Hohage [56] (cf. [30, Theorem 2.7] for Landweber iteration, [54] for the iteratively regularized Gauss–Newton method IRGNM, and [55] for generalized IRGNM).

Stopping Rules

In the context of ill-posed problems it is essential to stop iterative solution methods according to an appropriate rule to avoid an unbounded growth of the propagated noise. There are two possibilities, either a-priori rules or a-posteriori rules. A-priori rules [see, e.g., (58)] are computationally very effective. However, the disadvantage is that one has to know the smoothness index μ in (8) or (9) explicitly.

This is avoided in a-posteriori stopping rules. The most well-known a-posteriori criterion is the so-called *discrepancy principle*, i.e., the iteration is stopped after $k_* = k_*(\delta, y^\delta)$ steps with

$$\|y^\delta - F(x_{k_*}^\delta)\| \le \tau\delta < \|y^\delta - F(x_k^\delta)\|\,, \quad 0 \le k < k_*\,, \tag{10}$$

where $\tau > 1$.

The Lepskii type balancing principle is an interesting alternative a-posteriori rule, see [6, 8] and (62) below in the context of iterative methods.

If the noise level δ in (2) is unknown, heuristic stopping rules such as the quasioptimality principle, the Hanke–Raus rule, or the L-curve criterion still lead to convergence under certain structural assumptions on the noise, see [7, 74, 75, 89].

3 Gradient Methods

One way to derive iterative regularization methods is to apply gradient methods to the minimization problem

$$\min \tfrac{1}{2}\|F(x) - y\|^2 \quad \text{over} \quad \mathcal{D}(F)\,.$$

Since the negative gradient of this functional is given by $F'(x)^*(y - F(x))$ and taking into account that only noisy data y^δ are available, this yields methods of the form

$$x_{k+1}^\delta = x_k^\delta + \omega_k^\delta F'(x_k^\delta)^*(y^\delta - F(x_k^\delta))\,, \tag{11}$$

where $x_0^\delta = x_0$ is an initial guess of the exact solution. Choosing the factor ω_k^δ in a special way we obtain well-known methods like Landweber iteration, the steepest descent method, and the minimal error method.

Nonlinear Landweber Iteration

If one chooses $\omega_k^\delta = \omega$ to be constant, one obtains Landweber iteration. As already mentioned in the introduction of this chapter, well-definedness and convergence can only be proven if problem (1) is properly scaled. Without loss of generality we may assume that $\omega_k^\delta \equiv 1$ and that

$$\|F'(x)\| \le 1\,, \qquad x \in \mathcal{B}_{2\rho}(x_0) \subset \mathcal{D}(F)\,. \tag{12}$$

The nonlinear Landweber iteration is then given as the method

$$x_{k+1}^\delta = x_k^\delta + F'(x_k^\delta)^*(y^\delta - F(x_k^\delta))\,, \qquad k \in \mathbb{N}_0\,. \tag{13}$$

We want to emphasize that for fixed iteration index k the iterate x_k^δ depends continuously on the data y^δ, since x_k^δ is the result of a combination of continuous operations.

The results on convergence and convergence rates for this method presented here were established in [49] (see also [70]). To begin with, we formulate the following monotonicity property that gives us a clue how to choose the number τ in the stopping rule (10) (see [70, Proposition 2.2]).

Proposition 2. *Assume that the conditions (12) and (6) hold and that the equation $F(x) = y$ has a solution $x_* \in \mathcal{B}_\rho(x_0)$. If $x_k^\delta \in \mathcal{B}_\rho(x_*)$, a sufficient condition for x_{k+1}^δ to be a better approximation of x_* than x_k^δ is that*

$$\|y^\delta - F(x_k^\delta)\| > 2\,\frac{1+\eta}{1-2\eta}\,\delta\,.$$

Moreover, it then holds that $x_k^\delta, x_{k+1}^\delta \in \mathcal{B}_\rho(x_*) \subset \mathcal{B}_{2\rho}(x_0)$.

In view of this proposition, the number τ in the stopping rule (10) should be chosen as

$$\tau = 2\,\frac{1+\eta}{1-2\eta}$$

with η as in (6). To be able to prove that the stopping index k_* in (10) is finite and hence well defined it turns out that τ has to be chosen slightly larger (see [70, Corollary 2.3]), i.e.,

$$\tau > 2\,\frac{1+\eta}{1-2\eta} > 2\,. \tag{14}$$

Corollary 1. *Let the assumptions of Proposition 2 hold and let k_* be chosen according to the stopping rule (10), (14). Then*

$$k_*(\tau\delta)^2 < \sum_{k=0}^{k_*-1} \|y^\delta - F(x_k^\delta)\|^2 \le \frac{\tau}{(1-2\eta)\tau - 2(1+\eta)}\,\|x_0 - x_*\|^2.$$

In particular, if $y^\delta = y$ (i.e., if $\delta = 0$), then

$$\sum_{k=0}^{\infty} \|y - F(x_k)\|^2 < \infty\,. \tag{15}$$

Note that (15) implies that if Landweber iteration is run with precise data $y = y^\delta$, then the residual norms of the iterates tend to zero as $k \to \infty$. That is, if the iteration converges, then the limit is necessarily a solution of $F(x) = y$. The following convergence result holds (see [70, Theorem 2.4]):

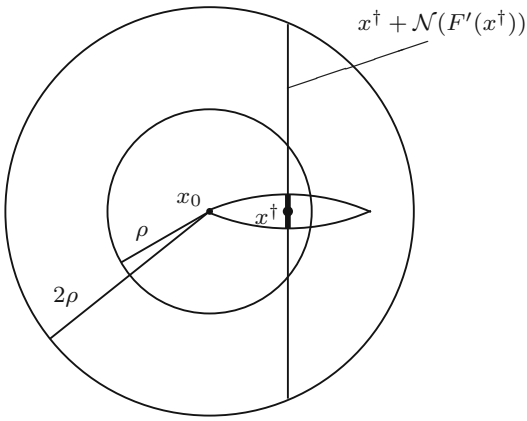


Fig. 1 The sketch shows the initial element x_0, the x_0-minimum-norm solution $x^\dagger$, the subset $x^\dagger + \mathcal{N}(F'(x^\dagger))$ and in bold the region, where the limit of the iterates x_k can be

Theorem 1. *Assume that the conditions* (12) *and* (6) *hold and that the equation* $F(x) = y$ *is solvable in* $\mathcal{B}_\rho(x_0)$. *Then the nonlinear Landweber iteration applied to exact data* y *converges to a solution of* $F(x) = y$. *If* $\mathcal{N}(F'(x^\dagger)) \subset \mathcal{N}(F'(x))$ *for all* $x \in \mathcal{B}_\rho(x^\dagger)$, *then* x_k *converges to* $x^\dagger$ *as* $k \to \infty$.

We emphasize that, in general, the limit of the Landweber iterates is no x_0-minimum-norm solution. However, since the monotonicity result of Proposition 2 holds for every solution, the limit of x_k has to be at least close to $x^\dagger$. As can be seen below, it has to be the closer the larger ρ can be chosen (Fig. 1).

It is well known that, if y^δ does not belong to the range of F, then the iterates x_k^δ of (13) cannot converge but still allow a stable approximation of a solution of $F(x) = y$ provided the iteration is stopped after k_* steps. The next result shows that the stopping rule (10), (14) renders the Landweber iteration a regularization method (see [70, Theorem 2.6]):

Theorem 2. *Let the assumptions of Theorem 1 hold and let* $k_* = k_*(\delta, y^\delta)$ *be chosen according to the stopping rule* (10), (14). *Then the Landweber iterates* $x_{k_*}^\delta$ *converge to a solution of* $F(x) = y$. *If* $\mathcal{N}(F'(x^\dagger)) \subset \mathcal{N}(F'(x))$ *for all* $x \in \mathcal{B}_\rho(x^\dagger)$, *then* $x_{k_*}^\delta$ *converges to* $x^\dagger$ *as* $\delta \to 0$.

To obtain convergence rates the exact solution has to satisfy some source conditions. Moreover, one has to guarantee that the iterates remain in $\mathcal{R}(F'(x^\dagger)^*)$. In [49] rates were proven under the additional assumption that F satisfies

$$F'(x) = R_x F'(x^\dagger) \quad \text{and} \quad \|R_x - I\| \le c\,\|x - x^\dagger\|, \qquad x \in \mathcal{B}_{2\rho}(x_0)\,,$$

where $\{R_x : x \in \mathcal{B}_{2\rho}(x_0)\}$ is a family of bounded linear operators $R_x : \mathcal{Y} \to \mathcal{Y}$ and c is a positive constant.

Unfortunately, these conditions are not always satisfied (see [49, Example 4.3]). Therefore, we consider instead of (13) the following slightly modified iteration method,

$$x_{k+1}^{\delta} = x_k^{\delta} + \omega G^{\delta}(x_k^{\delta})^*(y^{\delta} - F(x_k^{\delta}))\,, \quad k \in \mathbb{N}_0\,, \tag{16}$$

where, as above, $x_0^{\delta} = x_0$ is an initial guess, $G^{\delta}(x) := G(x, y^{\delta})$, and G is a continuous operator mapping $\mathcal{D}(F) \times \mathcal{Y}$ into $\mathcal{L}(\mathcal{X}, \mathcal{Y})$. The iteration will again be stopped according to the discrepancy principle (10).

To obtain local convergence and convergence rates for this modification we need the following assumptions:

Assumption 9. *Let ρ be a positive number such that $\mathcal{B}_{2\rho}(x_0) \subset \mathcal{D}(F)$.*

(i) The equation $F(x) = y$ has an x_0-minimum-norm solution $x^{\dagger}$ in $\mathcal{B}_{\rho}(x_0)$.
(ii) There exist positive constants c_1, c_2, c_3 and linear operators R_x^{δ} such that for all $x \in \mathcal{B}_{\rho}(x^{\dagger})$ the following estimates hold:

$$\|F(x) - F(x^{\dagger}) - F'(x^{\dagger})(x - x^{\dagger})\| \le c_1 \|F(x) - F(x^{\dagger})\| \, \|x - x^{\dagger}\|\,, \tag{17}$$

$$G^{\delta}(x) = R_x^{\delta} G^{\delta}(x^{\dagger})\,, \tag{18}$$

$$\|R_x^{\delta} - I\| \le c_2 \|x - x^{\dagger}\|\,, \tag{19}$$

$$\|F'(x^{\dagger}) - G^{\delta}(x^{\dagger})\| \le c_3 \delta\,. \tag{20}$$

(iii) The scaling parameter ω in (16) satisfies the condition

$$\omega \|F'(x^{\dagger})\|^2 \le 1\,.$$

Note that, if instead of (17) the slightly stronger condition

$$\begin{aligned} \|F(x) - F(\tilde{x}) - F'(x)(x - \tilde{x})\| &\le c\,\|x - \tilde{x}\| \\ \|F(x) - F(\tilde{x})\|, x, \tilde{x} &\in \mathcal{B}_{2\rho}(x_0) \subseteq \mathcal{D}(F), \end{aligned} \tag{21}$$

holds in $\mathcal{B}_{2\rho}(x_0)$ for some $c > 0$, then the unique existence of the x_0-minimum-norm solution $x^{\dagger}$ follows from Proposition 1 if $F(x) = y$ is solvable in $\mathcal{B}_{\rho}(x_0)$.

Convergence and convergence rates for the modification above are obtained as follows (see [70, Theorems 2.8 and 2.13]):

Theorem 3. *Let Assumption 9 hold and let $k_* = k_*(\delta, y^{\delta})$ be chosen according to the stopping rule (10).*

(i) If $\|x_0 - x^{\dagger}\|$ is so small and if the parameter τ in (10) is so large that

$$2\eta_1 + \eta_2^2 \eta_3^2 < 2$$

and

$$\tau > \frac{2(1+\eta_1+c_3\eta_2\|x_0-x^\dagger\|)}{2-2\eta_1-\eta_2^2\eta_3^2},$$

where

$$\begin{aligned}\eta_1 &:= \|x_0-x^\dagger\|(c_1+c_2(1+c_1\|x_0-x^\dagger\|)),\\ \eta_2 &:= 1+c_2\|x_0-x^\dagger\|,\\ \eta_3 &:= 1+2c_3\|x_0-x^\dagger\|,\end{aligned}$$

then the modified Landweber iterates $x_{k_*}^\delta$ converge to $x^\dagger$ as $\delta \to 0$.

(ii) If $\tau > 2$ and if $x^\dagger - x_0$ satisfies (8) with some $0 < \mu \le 1/2$ and $\|v\|$ sufficiently small, then it holds that

$$k_* = O\Big(\|v\|^{\frac{2}{2\mu+1}}\delta^{-\frac{2}{2\mu+1}}\Big)$$

and

$$\|x_{k_*}^\delta - x^\dagger\| = \begin{cases} o\Big(\|v\|^{\frac{1}{2\mu+1}}\delta^{\frac{2\mu}{2\mu+1}}\Big), & \mu < \frac{1}{2},\\ O\Big(\sqrt{\|v\|\delta}\Big), & \mu = \frac{1}{2}.\end{cases}$$

Note that for the modified Landweber iteration we obtain the same convergence rates and the same asymptotical estimate for k_* as for linear ill-posed problems (compare [36, Theorem 6.5]) if $\mu \le 1/2$ in (8).

Under the Assumption 9 and according to the theorem above the best possible convergence rate is

$$\|x_{k_*}^\delta - x^\dagger\| = O(\sqrt{\delta})$$

provided that $\mu = 1/2$. Even if $\mu > 1/2$ we cannot improve this rate without an additional restriction of the *nonlinearity* of F.

We will show for the following parameter estimation problem that the conditions of Assumption 9 are satisfied if $F'(x)$ is replaced by a certain operator $G^\delta(x)$.

Example 1. We treat the problem of estimating the diffusion coefficient a in

$$-(a(s)u(s)_s)_s = f(s), \quad s \in (0,1), \quad u(0) = 0 = u(1), \tag{22}$$

where $f \in L^2$; the subscript s denotes derivative with respect to s.

In this example, F is defined as the parameter-to-solution mapping

$$F : \mathcal{D}(F) := \{a \in H^1[0,1] \,:\, a(s) \geq \underline{a} > 0\} \to L^2[0,1]$$
$$a \mapsto F(a) := u(a)\,,$$

where $u(a)$ is the solution of (22). One can prove that F is Fréchet-differentiable (see, e.g., [24]) with

$$F'(a)h = A(a)^{-1}[(hu_s(a))_s]\,,$$
$$F'(a)^* w = -B^{-1}[u_s(a)(A(a)^{-1}w)_s]\,,$$

where

$$A(a) : H^2[0,1] \cap H_0^1[0,1] \to L^2[0,1]$$
$$u \mapsto A(a)u := -(au_s)_s$$

and

$$B : \mathcal{D}(B) := \{\psi \in H^2[0,1] \,:\, \psi'(0) = \psi'(1) = 0\} \to L^2[0,1]$$
$$\psi \mapsto B\psi := -\psi'' + \psi\,;$$

note that B^{-1} is the adjoint of the embedding operator from $H^1[0,1]$ in $L^2[0,1]$.

First of all, we show that F satisfies condition (17): let $F(a) = u$, $F(\tilde{a}) = \tilde{u}$, and $w \in L^2$. Noting that $(\tilde{u} - u) \in H^2 \cap H_0^1$ and that $A(a)$ is one-to-one and onto for $a, \tilde{a} \in \mathcal{D}(F)$ we obtain that

$$\begin{aligned}
&\langle\, F(\tilde{a}) - F(a) - F'(a)(\tilde{a} - a), w \,\rangle_{L^2} \\
&\qquad = \langle\, (\tilde{u} - u) - A(a)^{-1}[((\tilde{a} - a)u_s)_s], w \,\rangle_{L^2} \\
&\qquad = \langle\, A(a)(\tilde{u} - u) - ((\tilde{a} - a)u_s)_s, A(a)^{-1}w \,\rangle_{L^2} \\
&\qquad = \langle\, ((\tilde{a} - a)(\tilde{u}_s - u_s))_s, A(a)^{-1}w \,\rangle_{L^2} \\
&\qquad = -\langle\, (\tilde{a} - a)(\tilde{u} - u)_s, (A(a)^{-1}w)_s \,\rangle_{L^2} \\
&\qquad = \langle\, F(\tilde{a}) - F(a), ((\tilde{a} - a)(A(a)^{-1}w)_s)_s \,\rangle_{L^2}\,.
\end{aligned}$$

This together with the fact that $\|g\|_{L^\infty} \leq \sqrt{2}\|g\|_{H^1}$ and that $\|g\|_{L^\infty} \leq \|g'\|_{L^2}$ if $g \in H^1$ is such that $g(\xi) = 0$ for some $\xi \in [0,1]$ yields the estimate

$$\begin{aligned}
&\|F(\tilde{a}) - F(a) - F'(a)(\tilde{a} - a)\|_{L^2} \\
&\quad \leq \sup_{\|w\|_{L^2}=1} \langle\, F(\tilde{a}) - F(a), ((\tilde{a} - a)(A(a)^{-1}w)_s)_s \,\rangle_{L^2}
\end{aligned}$$

$$\leq \|F(\tilde{a}) - F(a)\|_{L^2} \sup_{\|w\|_{L^2}=1} \left[\left\| \left(\frac{\tilde{a}-a}{a} \right)_s \right\|_{L^2} \|a(A(a)^{-1}w)_s\|_{L^\infty} + \left\| \frac{\tilde{a}-a}{a} \right\|_{L^\infty} \|w\|_{L^2} \right]$$

$$\leq \underline{a}^{-1}(1 + \sqrt{2} + \underline{a}^{-1}\sqrt{2}\|a\|_{H^1}) \|F(\tilde{a}) - F(a)\|_{L^2} \|\tilde{a} - a\|_{H^1} . \tag{23}$$

This implies (17).

The conditions (18) and (19) are not fulfilled with $G^\delta(x) = F'(x)$. Noting that $F'(a)^* w$ is the unique solution of the variational problem: for all $v \in H^1$

$$\langle (F'(a)^* w)_s, v_s \rangle_{L^2} + \langle F'(a)^* w, v \rangle_{L^2} = \langle u(a), ((A(a)^{-1}w)_s v)_s \rangle_{L^2} , \tag{24}$$

we propose to choose G^δ in (16) as follows: $G^\delta(a)^* w = G(a, u^\delta)^* w$ is the unique solution g of the variational problem

$$\langle g_s, v_s \rangle_{L^2} + \langle g, v \rangle_{L^2} = \langle u^\delta, ((A(a)^{-1}w)_s v)_s \rangle_{L^2} , \quad v \in H^1 . \tag{25}$$

This operator G^δ obviously satisfies (18), since

$$G(\tilde{a}, u^\delta)^* = G(a, u^\delta)^* R(\tilde{a}, a)^*$$

with

$$R(\tilde{a}, a)^* = A(a)A(\tilde{a})^{-1}.$$

The condition (19) is satisfied, since one can estimate as in (23) that

$$\|R(\tilde{a}, a)^* - I\| = \|A(a)A(\tilde{a})^{-1} - I\| \leq \underline{a}^{-1}(1 + \sqrt{2} + \underline{a}^{-1}\sqrt{2}\|\tilde{a}\|_{H^1}) \|\tilde{a} - a\|_{H^1} .$$

Note that a constant c_2 independent from $\tilde{a}$ can be found, since it is assumed that $\tilde{a} \in \mathcal{B}_\rho(a)$. Now we turn to condition (20): using (24) and (25) we obtain similarly to (23) the estimate

$$\|(F'(a)^* - G(a, u^\delta)^*)w\|_{H^1} = \sup_{\|v\|_{H^1}=1} \langle u(a) - u^\delta, ((A(a)^{-1}w)_s v)_s \rangle_{L^2}$$

$$\leq \underline{a}^{-1}(1 + \sqrt{2} + \underline{a}^{-1}\sqrt{2}\|a\|_{H^1}) \|u(a) - u^\delta\|_{L^2} \|w\|_{L^2} .$$

This together with $F(a^\dagger) = u(a^\dagger)$ and $\|u^\delta - u(a^\dagger)\|_{L^2} \leq \delta$ implies that

$$\|F'(a^\dagger) - G(a^\dagger, u^\delta)\| \leq \underline{a}^{-1}(1 + \sqrt{2} + \underline{a}^{-1}\sqrt{2}\|a^\dagger\|_{H^1}) \delta$$

and hence (20) holds.

Thus, Theorem 3 is applicable, i.e., if ω and τ are chosen appropriately, then the modified Landweber iterates $a_{k_*}^\delta$ (cf. (16)) where k_* is chosen according to the stopping rule (10) converge to the exact solution $a^\dagger$ with the rate $O(\sqrt{\delta})$ provided that

$$a^\dagger - a_0 = -B^{-1}[u_s(a^\dagger)(A(a^\dagger)^{-1}w)_s]$$

with $\|w\|$ sufficiently small. Note that this means that

$$a^\dagger - a_o \in H^3\,, \qquad (a^\dagger - a_0)_s(0) = 0 = (a^\dagger - a_0)_s(1)\,,$$
$$z := \frac{(a^\dagger - a_0)_{ss} - (a^\dagger - a_0)}{u_s(a)} \in H^1\,, \qquad \int_0^1 z(s)\,ds = 0\,.$$

Basically this means that one has to know all rough parts of $a^\dagger$ up to H^3. But without this knowledge one cannot expect to get the rate $O(\sqrt{\delta})$.

In [49] two other nonlinear problems were treated where conditions (18) and (19) are satisfied with $G^\delta(x) = F'(x)$.

Landweber Iteration in Hilbert Scales

We have mentioned in the last subsection that for classical Landweber iteration the rates cannot be better than $O(\sqrt{\delta})$ under the given assumptions. However, better rates may be obtained for solutions that satisfy stronger smoothness conditions if the iteration is performed in a subspace of $\mathcal{X}$ with a stronger norm. This leads us directly to regularization in Hilbert scales. On the other hand for solutions with poor smoothness properties the number of iterations may be reduced if the iteration is performed in a space with a weaker norm.

First of all, we shortly repeat the definition of a Hilbert scale: let L be a densely defined unbounded self-adjoint strictly positive operator in $\mathcal{X}$. Then $(\mathcal{X}_s)_{s\in\mathbb{R}}$ denotes the Hilbert scale induced by L if $\mathcal{X}_s$ is the completion of $\bigcap_{k=0}^\infty D(L^k)$ with respect to the Hilbert space norm $\|x\|_s := \|L^s x\|_{\mathcal{X}}$; obviously, $\|x\|_0 = \|x\|_{\mathcal{X}}$ (see [78] or [36, Section 8.4] for details).

The operator $F'(x_k^\delta)^*$ in (13) will now be replaced by the adjoint of $F'(x_k^\delta)$ considered as an operator from $\mathcal{X}_s$ into $\mathcal{Y}$. Usually $s \geq 0$, but we will see below that there are special cases where a negative choice of s can be advantageous. Since by definition of $\mathcal{X}_s$ this adjoint is given by $L^{-2s}F'(x_k^\delta)^*$, (13) is replaced by the iteration process

$$x_{k+1}^\delta = x_k^\delta + L^{-2s}F'(x_k^\delta)^*(y^\delta - F(x_k^\delta))\,, \qquad k \in \mathbb{N}_0\,. \tag{26}$$

As in the previous chapter the iteration process is stopped according to the discrepancy principle (10).

Proofs of convergence and convergence rates for this method can be found in [34, 70, 88]. For an approach, where the Hilbert scale is chosen in the space Y, see [33].

The following basic conditions are needed.

Assumption 10.

(i) $F : \mathcal{D}(F)(\subset \mathcal{X}) \to \mathcal{Y}$ *is continuous and Fréchet-differentiable in* $\mathcal{X}$.
(ii) $F(x) = y$ *has a solution* $x^\dagger$.
(iii) $\|F'(x^\dagger)x\| \leq \overline{m}\|x\|_{-a}$ *for all* $x \in \mathcal{X}$ *and some* $a > 0$, $\overline{m} > 0$. *Moreover, the extension of* $F'(x^\dagger)$ *to* $\mathcal{X}_{-a}$ *is injective.*
(iv) $B := F'(x^\dagger)L^{-s}$ *is such that* $\|B\|_{\mathcal{X},\mathcal{Y}} \leq 1$, *where* $-a < s$. *If* $s < 0$, $F'(x^\dagger)$ *has to be replaced by its extension to* $\mathcal{X}_s$.

Usually, for the analysis of regularization methods in Hilbert scales a stronger condition than (iii) is used, namely (cf., e.g., [88])

$$\|F'(x^\dagger)x\| \sim \|x\|_{-a} \qquad \text{for all } x \in \mathcal{X}, \tag{27}$$

where the number a can be interpreted as a *degree of ill-posedness* of the linearized problem in $x^\dagger$. However, this condition is not always fulfilled. Sometimes one can only prove that condition (iii) in Assumption 10 holds. It might also be possible that one can prove an estimate from below in a slightly weaker norm (see examples in [34]), i.e.,

$$\|F'(x^\dagger)x\| \geq \underline{m}\|x\|_{-\tilde{a}} \qquad \text{for all } x \in \mathcal{X} \text{ and some } \tilde{a} \geq a, \underline{m} > 0. \tag{28}$$

The next proposition sheds more light onto condition (iii) in Assumption 10 and (28). The proof follows the lines of [36, Corollary 8.22] noting that the results there not only hold for $s \geq 0$ but also for $s > -a$.

Proposition 3. *Let Assumption 10 hold. Then for all* $\nu \in [0, 1]$ *it holds that*

$$\begin{aligned} \mathcal{D}((B^*B)^{-\frac{\nu}{2}}) = \mathcal{R}((B^*B)^{\frac{\nu}{2}}) &\subset \mathcal{X}_{\nu(a+s)}, \\ \|(B^*B)^{\frac{\nu}{2}}x\| &\leq \overline{m}^\nu \|x\|_{-\nu(a+s)} \qquad \text{for all } x \in \mathcal{X}, \\ \|(B^*B)^{-\frac{\nu}{2}}x\| &\geq \overline{m}^{-\nu} \|x\|_{\nu(a+s)} \qquad \text{for all } x \in \mathcal{D}((B^*B)^{-\frac{\nu}{2}}). \end{aligned}$$

Note that condition (iii) is equivalent to

$$\mathcal{R}(F'(x^\dagger)^*) \subset \mathcal{X}_a \quad \text{and} \quad \|F'(x^\dagger)^*w\|_a \leq \overline{m}\|w\| \quad \text{for all} \quad w \in \mathcal{Y}.$$

If in addition condition (28) holds, then for all $\nu \in [0, 1]$ *it holds that*

$$\mathcal{X}_{\nu(\tilde{a}+s)} \subset \mathcal{R}((B^*B)^{\frac{\nu}{2}}) = \mathcal{D}((B^*B)^{-\frac{\nu}{2}}),$$
$$\|(B^*B)^{\frac{\nu}{2}}x\| \geq \underline{m}^{\nu}\|x\|_{-\nu(\tilde{a}+s)} \qquad \text{for all} \quad x \in \mathcal{X},$$
$$\|(B^*B)^{-\frac{\nu}{2}}x\| \leq \underline{m}^{-\nu}\|x\|_{\nu(\tilde{a}+s)} \qquad \text{for all} \quad x \in \mathcal{X}_{\nu(\tilde{a}+s)}.$$

Note that condition (28) is equivalent to

$$\mathcal{X}_{\tilde{a}} \subset \mathcal{R}(F'(x^\dagger)^*) \text{ and } \|F'(x^\dagger)^*w\|_{\tilde{a}} \geq \underline{m}\|w\| \quad \text{for all} \quad w \in \mathcal{N}(F'(x^\dagger)^*)^\perp \text{ with } F'(x^\dagger)^*w \in \mathcal{X}_{\tilde{a}}.$$

In our convergence analysis the following *shifted* Hilbert scale will play an important role

$$\tilde{\mathcal{X}}_r := \mathcal{D}((B^*B)^{\frac{s-r}{2(a+s)}}L^s) \quad \text{equipped with the norm}$$
$$|||x|||_r := \|(B^*B)^{\frac{s-r}{2(a+s)}}L^s x\|_{\mathcal{X}},$$

where a, s, and B are as in Assumption 10. Some properties of this shifted Hilbert scale can be found in [70, Proposition 3.3].

For the convergence rates analysis we need the following smoothness conditions on the solution $x^\dagger$ and the Fréchet-derivative of F.

Assumption 11.

(i) $x_0 \in \tilde{\mathcal{B}}_\rho(x^\dagger) := \{x \in \mathcal{X} : x - x^\dagger \in \tilde{\mathcal{X}}_0 \wedge |||x - x^\dagger|||_0 \leq \rho\} \subset \mathcal{D}(F)$ *for some* $\rho > 0$.

(ii) $\|F'(x^\dagger) - F'(x)\|_{\tilde{\mathcal{X}}_{-b},\mathcal{Y}} \leq c\,|||x^\dagger - x|||_0^\beta$ *for all* $x \in \tilde{\mathcal{B}}_\rho(x^\dagger)$ *and some* $b \in [0, a]$, $\beta \in (0, 1]$, *and* $c > 0$.

(iii) $x^\dagger - x_0 \in \tilde{\mathcal{X}}_u$ *for some* $(a-b)/\beta < u \leq b + 2s$, *i.e., there is an element* $v \in \mathcal{X}$ *so that*

$$L^s(x^\dagger - x_0) = (B^*B)^{\frac{u-s}{2(a+s)}}v \quad \text{and} \quad \|v\|_0 = |||x_0 - x^\dagger|||_u.$$

Condition (iii) is a smoothness condition for the exact solution comparable to (8). Usually $\mathcal{X}_u$ is used instead of $\tilde{\mathcal{X}}_u$. However, these conditions are equivalent if (27) holds.

For the proof of the next convergence rates result see [70, Theorem 3.8].

Theorem 4. *Let Assumptions 10 and 11 hold. Moreover, let* $k_* = k_*(\delta, y^\delta)$ *be chosen according to the stopping rule (10) with* $\tau > 2$ *and let* $|||x_0 - x^\dagger|||_u$ *be sufficiently small. Then the following estimates are valid for* $\delta > 0$ *and some positive constants* c_r*:*

$$k_* \leq \left(\tfrac{2\tau}{\tau-2}\,|||x_0 - x^\dagger|||_u\,\delta^{-1}\right)^{\frac{2(a+s)}{a+u}} \tag{29}$$

and for $-a \le r < u$

$$\|\!| x_{k_*}^\delta - x^\dagger |\!\|_r \le c_r \, \|\!| x_0 - x^\dagger |\!\|_u^{\frac{a+r}{a+u}} \, \delta^{\frac{u-r}{a+u}} .$$

As usual for regularization in Hilbert scales, we are interested in obtaining convergence rates with respect to the norm in $\mathcal{X} = \mathcal{X}_0$.

Corollary 2. *Under the assumptions of Theorem 4 the following estimates hold:*

$$\|x_{k_*}^\delta - x^\dagger\| = O\left(\delta^{\frac{u}{a+u}}\right) \qquad \text{if} \quad s \le 0 , \tag{30}$$

$$\|x_{k_*}^\delta - x^\dagger\| = O\left(\|x_{k_*}^\delta - x^\dagger\|_s\right) = O\left(\delta^{\frac{u-s}{a+u}}\right) \qquad \text{if} \quad 0 < s < u .$$

If in addition (28) *holds, then for* $s > 0$ *the rate can be improved to*

$$\|x_{k_*}^\delta - x^\dagger\| = O\left(\|\!| x_{k_*}^\delta - x^\dagger |\!\|_r\right) = O\left(\delta^{\frac{u-r}{a+u}}\right) \quad \text{if} \quad r := \tfrac{s(\tilde{a}-a)}{\tilde{a}+s} \le u .$$

Note that (29) implies that k_* is finite for $\delta > 0$ and hence $x_{k_*}^\delta$ is a stable approximation of $x^\dagger$.

Moreover, it can be seen from (29) that the larger s the faster k_* possibly grows if $\delta \to 0$. As a consequence, s should be kept as small as possible to reduce the number of iterations and hence to reduce the numerical effort. If u is close to 0, it might be possible to choose a negative s. According to (30), we would still get the optimal rate, but, due to (29), k_* would not grow so fast. Choosing a negative s could be interpreted as a preconditioned Landweber method (cf. [34]).

We will now comment on the rates in Corollary 2: if only Assumption 10 (iii) is satisfied, i.e., if $\|F'(x^\dagger)x\|$ may be estimated through the norm in $\mathcal{X}_{-a}$ only from above, convergence rates in $\mathcal{X}$ can only be given if $s < u$, i.e., only for the case of undersmoothing. If $s > 0$, the rates will not be optimal in general. To obtain rates also for $s > u$, i.e., for the case of oversmoothing, condition (28) has to be additionally satisfied. From what we said on the choice of s above, the case of oversmoothing is not desirable. However, note that the rates for $\|x_{k_*}^\delta - x^\dagger\|_0$ can be improved if (28) holds also for $0 < s < u$. Moreover, if $\tilde{a} = a$, i.e., if the usual equivalence condition (27) is satisfied, then we always obtain the usual optimal rates $O(\delta^{\frac{u}{a+u}})$ (see [87]).

For numerical computations one has to approximate the infinite-dimensional spaces by finite-dimensional ones. Also the operators F and $F'(x)^*$ have to be approximated by suitable finite-dimensional realizations. An appropriate convergence rates analysis has been carried out in [88]. This analysis also shows that a modification, where $F'(x_k^\delta)^*$ in (26) is replaced by $G^\delta(x_k^\delta)$ similar as in (16), is possible.

Steepest Descent and Minimal Error Method

These two methods are again of the form (11), where the coefficients ω_k^δ are chosen as

$$\omega_k^\delta := \frac{\|s_k^\delta\|^2}{\|F'(x_k^\delta)s_k^\delta\|^2} \qquad \text{and} \qquad \omega_k^\delta := \frac{\|y^\delta - F(x_k^\delta)\|^2}{\|s_k^\delta\|^2}$$

for the *steepest descent method* and for the *minimal error method*, respectively.

In [35] it has been shown that even for the solution of *linear* ill-posed problems the steepest descent method is only a regularization method when stopped via a discrepancy principle and not via an a-priori parameter choice strategy. Therefore, we will use (10), (14) as stopping rule.

Again one can show the monotonicity of the errors and well-definedness of the steepest descent and minimal error method (see [70, Proposition 3.20]). Convergence can be shown for perturbed data (see, e.g., [70, Theorem 3.22]). However, so far, convergence rates were proved only in the case of exact data (see [90]).

Further Literature on Gradient Methods

Iteratively Regularized Landweber Iteration

By adding an additional penalty term to the iteration scheme of classical Landweber iteration, i.e.,

$$x_{k+1}^\delta = x_k^\delta + F'(x_k^\delta)^*(y^\delta - F(x_k^\delta)) + \beta_k(x_0 - x_k^\delta) \quad \text{with} \quad 0 < \beta_k \le \beta_{\max} < \tfrac{1}{2}\,.$$

one can obtain convergence rates results under weaker restrictions on the nonlinearity of F (see [70, Section 3.2], [98]). The additional term is motivated by the iteratively regularized Gauss–Newton method, see section "Iteratively Regularized Gauss–Newton Method".

A Derivative Free Approach

Based on an idea by Engl and Zou [37], Kügler, in his thesis [80] (see also [79]), developed a modification of Landweber iteration for parameter identification problems where it is not needed that F is Fréchet-differentiable.

Generalization to Banach Spaces

A generalization of Landweber iteration to the case where $\mathcal{X}$ and $\mathcal{Y}$ are Banach spaces was considered in the papers [71,99,100], see also the book [101]. The basic version (in case of reflexive preimage space X) reads as

$$x_{k+1}^\delta = j_{q^*}^{X^*}\left(j_q^X(x_k^\delta) + \omega_k^\delta F'(x_k^\delta)^* j_p^Y(y^\delta - F(x_k^\delta))\right),$$

where $p, q \in (1, \infty)$, $q^* = \frac{q}{q-1}$, X^* is the dual of X, j_q^X denotes a single valued selection from the set valued duality mapping $J_q^X = \partial\left(\frac{1}{q}\|\cdot\|^q\right)$, $J_q^X : X \to 2^{X^*}$, and ω_k^δ is an appropriately chosen step size. Banach space versions of the iteratively regularized Landweber iteration can be found in [51, 64].

4 Newton Type Methods

Newton's method for the nonlinear operator equation (1) reads as

$$F'(x_k^\delta)(x_{k+1}^\delta - x_k^\delta) = y^\delta - F(x_k^\delta)\,. \tag{31}$$

Since ill-posedness of the nonlinear problem (1) is usually inherited by its linearization (31), regularization has to be applied in each Newton step. Formulating (31) as a least squares problem

$$\min_{x \in \mathcal{D}(F)} \|y^\delta - F(x_k^\delta) - F'(x_k^\delta)(x - x_k^\delta)\|^2$$

and applying Tikhonov regularization leads to either the Levenberg–Marquardt method

$$x_{k+1}^\delta = \arg\min_{x \in \mathcal{D}(F)} \|y^\delta - F(x_k^\delta) - F'(x_k^\delta)(x - x_k^\delta)\|^2 + \alpha_k \|x - x_k^\delta\|^2\,, \tag{32}$$

where the regularization term $\|x - x_k^\delta\|^2$ is updated in each Newton step, or the iteratively regularized Gauss–Newton method (IRGNM)

$$x_{k+1}^\delta = \arg\min_{x \in \mathcal{D}(F)} \|y^\delta - F(x_k^\delta) - F'(x_k^\delta)(x - x_k^\delta)\|^2 + \alpha_k \|x - x_0\|^2 \tag{33}$$

with a fixed a-priori guess $x_0 \in \mathcal{X}$. The choice of the sequence of regularization parameters α_k and the main ideas of the convergence analysis are quite different for both methods as will be outlined in the following subsections.

Levenberg–Marquardt and Inexact Newton Methods

In the Hilbert space setting with an open set $\mathcal{D}(F)$, the minimizer of the quadratic functional in (32) can be written as the solution of a linear system which leads to the formulation

$$x_{k+1}^\delta = x_k^\delta + (F'(x_k^\delta)^* F'(x_k^\delta) + \alpha_k I)^{-1} F'(x_k^\delta)^*(y^\delta - F(x_k^\delta))\,, \tag{34}$$

of the Levenberg–Marquardt method that can as well be motivated by a trust region approach.

Our exposition follows the seminal paper by Hanke [46], in which the first convergence analysis for this class of Newton type methods was given. According to this paper, α_k should be chosen such that

$$\|y^\delta - F(x_k^\delta) - F'(x_k^\delta)(x_{k+1}^\delta(\alpha_k) - x_k^\delta)\| = q\,\|y^\delta - F(x_k^\delta)\| \tag{35}$$

for some $q \in (0,1)$, where $x_{k+1}^\delta(\alpha)$ is defined as in (34) with α_k replaced by α. This means that the Newton equation (31) is only solved up to a residual of magnitude $q\,\|y^\delta - F(x_k^\delta)\|$ which corresponds to the concept of inexact Newton methods as they were first considered for well-posed problems in [28]. It can be shown (see [46]) that (35) has a unique solution α_k provided that

$$\|y^\delta - F(x_k^\delta) - F'(x_k^\delta)(x^\dagger - x_k^\delta)\| \le \tfrac{q}{\gamma}\,\|y^\delta - F(x_k^\delta)\| \tag{36}$$

for some $\gamma > 1$ which in its turn can be guaranteed by (21).

The key step in the convergence analysis of the Levenberg–Marquardt method is to show monotonicity of the error norms $\|x_k^\delta - x^\dagger\|$. To sketch this monotonicity proof we assume that (36) holds and therewith the parameter choice (35) is feasible. Using the notation $K_k = F'(x_k^\delta)$ as well as Cauchy–Schwarz inequality and the identity

$$\alpha_k (K_k K_k^* + \alpha_k I)^{-1}(y^\delta - F(x_k^\delta)) = y^\delta - F(x_k^\delta) - K_k(x_{k+1}^\delta - x_k^\delta)\,,$$

we get

$$\begin{aligned}
&\|x_{k+1}^\delta - x^\dagger\|^2 - \|x_k^\delta - x^\dagger\|^2 \\
&\quad = 2\langle\, x_{k+1}^\delta - x_k^\delta, x_k^\delta - x^\dagger\,\rangle + \|x_{k+1}^\delta - x_k^\delta\|^2 \\
&\quad = \langle\, (K_k K_k^* + \alpha_k I)^{-1}(y^\delta - F(x_k^\delta)), \\
&\qquad\quad 2K_k(x_k^\delta - x^\dagger) + (K_k K_k^* + \alpha_k I)^{-1} K_k K_k^*(y^\delta - F(x_k^\delta))\,\rangle \\
&\quad = -2\alpha_k\,\|(K_k K_k^* + \alpha_k I)^{-1}(y^\delta - F(x_k^\delta))\|^2 \\
&\qquad - \|(K_k^* K_k + \alpha_k I)^{-1} K_k^*(y^\delta - F(x_k^\delta))\|^2 \\
&\qquad + 2\langle\, (K_k K_k^* + \alpha_k I)^{-1}(y^\delta - F(x_k^\delta)),\, y^\delta - F(x_k^\delta) - K_k(x^\dagger - x_k^\delta)\,\rangle \\
&\quad \le -\|x_{k+1}^\delta - x_k^\delta\|^2 - 2\alpha_k^{-1}\,\|y^\delta - F(x_k^\delta) - K_k(x_{k+1}^\delta - x_k^\delta)\| \cdot \\
&\qquad \Big(\|y^\delta - F(x_k^\delta) - K_k(x_{k+1}^\delta - x_k^\delta)\| - \|y^\delta - F(x_k^\delta) - K_k(x^\dagger - x_k^\delta)\|\Big)\,.
\end{aligned} \tag{37}$$

By (36) and the parameter choice (35), we have

$$\|y^\delta - F(x_k^\delta) - K_k(x^\dagger - x_k^\delta)\| \le \gamma^{-1}\,\|y^\delta - F(x_k^\delta) - K_k(x_{k+1}^\delta - x_k^\delta)\|\,.$$

Thus, (37) and $\gamma > 1$ imply estimates (38) and (39) in the following proposition (see [70, Proposition 4.1]):

Proposition 4. *Let $0 < q < 1 < \gamma$ and assume that (1) has a solution and that (36) holds so that α_k can be defined via (35). Then, the following estimates hold:*

$$\|x_k^\delta - x^\dagger\|^2 - \|x_{k+1}^\delta - x^\dagger\|^2 \geq \|x_{k+1}^\delta - x_k^\delta\|^2 , \tag{38}$$

$$\begin{aligned} &\|x_k^\delta - x^\dagger\|^2 - \|x_{k+1}^\delta - x^\dagger\|^2 \\ &\quad \geq \frac{2(\gamma-1)}{\gamma\alpha_k} \|y^\delta - F(x_k^\delta) - F'(x_k^\delta)(x_{k+1}^\delta - x_k^\delta)\|^2 \qquad (39) \\ &\quad \geq \frac{2(\gamma-1)(1-q)q}{\gamma \|F'(x_k^\delta)\|^2} \|y^\delta - F(x_k^\delta)\|^2 . \qquad (40) \end{aligned}$$

Based on the resulting weak convergence of a subsequence of x_k^δ as well as on quadratic summability of the (linearized) residuals, which can be easily obtained by summing up both sides of (39) and (40), one obtains convergence as $k \to \infty$ in case of exact data [70, Theorem 4.2]:

Theorem 5. *Let $0 < q < 1$ and assume that (1) is solvable in $\mathcal{B}_\rho(x_0)$, that F' is uniformly bounded in $\mathcal{B}_\rho(x^\dagger)$, and that the Taylor remainder of F satisfies (21) for some $c > 0$. Then the Levenberg–Marquardt method with exact data $y^\delta = y$, $\|x_0 - x^\dagger\| < q/c$ and α_k determined from (35), converges to a solution of $F(x) = y$ as $k \to \infty$.*

In case of noisy data, Hanke [46] proposes to stop the iteration according to the discrepancy principle (10) and proves convergence as $\delta \to 0$ (see, e.g., [70, Theorem 4.3]):

Theorem 6. *Let the assumptions of Theorem 5 hold. Additionally let $k_* = k_*(\delta, y^\delta)$ be chosen according to the stopping rule (10) with $\tau > 1/q$. Then for $\|x_0 - x^\dagger\|$ sufficiently small, the discrepancy principle (10) terminates the Levenberg–Marquardt method with α_k determined from (35) after finitely many iterations k_*, and*

$$k_*(\delta, y^\delta) = O(1 + |\ln \delta|) .$$

Moreover, the Levenberg–Marquardt iterates $x_{k_*}^\delta$ converge to a solution of $F(x) = y$ as $\delta \to 0$.

Convergence rates seem to be much harder to prove for the Levenberg–Marquardt method than for the iteratively regularized Gauss–Newton method (see section "Iteratively Regularized Gauss–Newton Method"). Suboptimal rates under source conditions (8) have been proven by Rieder [94, 95] under the nonlinearity assumption

$$F'(x) = R_x F'(x^\dagger) \quad \text{and} \quad \|I - R_x\| \le c_R \|x - x^\dagger\|\,, \qquad x \in \mathcal{B}_\rho(x_0) \subseteq \mathcal{D}(F)\,, \tag{41}$$

where c_R is a positive constant. Only quite recently, Hanke [48, Theorem 2.1] proved the following optimal rates result:

Theorem 7. *Let a solution $x^\dagger$ of (1) exist and let (41) as well as (8) hold with some $0 < \mu \le 1/2$ and $\|v\|$ sufficiently small. Moreover, let α_k and k_* be chosen according to (35) and (10), respectively with $\tau > 2$ and $1 > q > 1/\tau$. Then the Levenberg–Marquardt iterates defined by (34) remain in $\mathcal{B}_\rho(x_0)$ and converge with the rate*

$$\|x_{k_*}^\delta - x^\dagger\| = O\left(\delta^{\frac{2\mu}{2\mu+1}}\right).$$

Finally, we quote the rates result [70, Theorem 4.7] that is almost optimal and instead of the a-posteriori choices of α_k and k_*, presumes a geometrically decreasing sequence of regularization parameters, i.e.,

$$\alpha_k = \alpha_0 q^k\,, \qquad \text{for some} \quad \alpha_0 > 0\,, \quad q \in (0,1)\,, \tag{42}$$

and the following a-priori stopping rule

$$\eta_{k_*}\alpha_{k_*}^{\mu+\frac{1}{2}} \le \delta < \eta_k \alpha_k^{\mu+\frac{1}{2}}\,, \quad 0 \le k < k_*, \quad \eta_k := \eta(k+1)^{-(1+\varepsilon)}\,, \quad \text{for some} \quad \eta > 0\,, \quad \varepsilon > 0\,. \tag{43}$$

Theorem 8. *Let a solution $x^\dagger$ of (1) exist and let (41) as well as (8) hold with some $0 < \mu \le 1/2$ and $\|v\|$ sufficiently small. Moreover, let α_k and k_* be chosen according to (42) and (43) with η sufficiently small, respectively. Then the Levenberg–Marquardt iterates defined by (34) remain in $\mathcal{B}_\rho(x_0)$ and converge with the rate*

$$\|x_{k_*}^\delta - x^\dagger\| = O\left((\delta\,(1 + |\ln\delta|)^{(1+\varepsilon)})^{\frac{2\mu}{2\mu+1}}\right).$$

Moreover,

$$\|F(x_{k_*}^\delta) - y\| = O\left(\delta\,(1 + |\ln\delta|)^{(1+\varepsilon)}\right)$$

and

$$k_* = O(1 + |\ln\delta|)\,.$$

For the noise free case ($\delta = 0$, $\eta = 0$) we obtain that

$$\|x_k - x^\dagger\| = O(\alpha_k^\mu)\,,$$

and that

$$\|F(x_k) - y\| = O\left(\alpha_k^{\mu+\frac{1}{2}}\right).$$

Further Literature on Inexact Newton Methods

Hanke [47] and Rieder [94–96] have extended the Levenberg–Marquardt method by proposing regularization methods other than Tikhonov in the inexact solution of the Newton equation

$$x_{k+1}^\delta = x_k^\delta + \Phi(F'(x_k^\delta), y^\delta - F(x_k^\delta)),$$

with $\Phi(F'(x_k^\delta), y^\delta - F(x_k^\delta))$, e.g., defined by the conjugate gradient method.

Recently, Hochbruck et al. [53] proposed the application of an exponential Euler scheme to the Showalter differential equation

$$x'(t) = F'(x(t))^*(y^\delta - F(x_k^\delta)),$$

which leads to a Newton type iterative method of the form

$$x_{k+1}^\delta = x_k^\delta + h_k \phi(-h_k\, F'(x_k^\delta)^* F'(x_k^\delta)) F'(x_k^\delta)^*(y^\delta - F(x_k^\delta)),$$

with

$$\phi(z) = \frac{e^z - 1}{z}.$$

In [52] they show convergence using the discrepancy principle (10) as a stopping rule under condition (6), as well as optimal convergence rates under the condition that

$$F'(x) = R_x F'(x^\dagger) \quad \text{and} \quad \|I - R_x\| \le c_R, \qquad x \in \mathcal{B}_\rho(x^\dagger) \subseteq \mathcal{D}(F), \tag{44}$$

for some $c_R \in (0,1)$, and under the source condition (8) with $\mu \le 1/2$ for an appropriate choice of the pseudo time step size h_k.

Iteratively Regularized Gauss–Newton Method

In the Hilbert space setting, the variational formulation (33) of the iteratively regularized Gauss–Newton method can be equivalently written as

$$x_{k+1}^\delta = x_k^\delta + (F'(x_k^\delta)^* F'(x_k^\delta) + \alpha_k I)^{-1}(F'(x_k^\delta)^*(y^\delta - F(x_k^\delta)) + \alpha_k(x_0 - x_k^\delta)). \tag{45}$$

Here the sequence of regularization parameters is a-priori chosen such that

$$\alpha_k > 0\,, \qquad 1 \le \frac{\alpha_k}{\alpha_{k+1}} \le r\,, \qquad \lim_{k\to\infty} \alpha_k = 0\,, \tag{46}$$

for some $r > 1$.

This method was first proposed and analyzed by Bakushinskii [3], see also [5] and the references therein, as well as [15, 54, 66, 69, 70]. The results presented here and in section "Generalizations of the IRGNM" together with proofs and further details can be found in [70].

The key point in the convergence analysis of the iteratively regularized Gauss–Newton method is the fact that under a source condition (8) the error $\|x_{k+1}^\delta - x^\dagger\|$ is up to some *small* additional terms equal to $\alpha_k^\mu w_k(\mu)$ with $w_k(s)$ defined as in the following lemma that is easy to prove.

Lemma 1. *Let $K \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$, $s \in [0,1]$, and let $\{\alpha_k\}$ be a sequence satisfying $\alpha_k > 0$ and $\alpha_k \to 0$ as $k \to \infty$. Then it holds that*

$$w_k(s) := \alpha_k^{1-s} \|(K^*K + \alpha_k I)^{-1}(K^*K)^s v\| \le s^s(1-s)^{1-s}\|v\| \le \|v\| \tag{47}$$

and that

$$\lim_{k\to\infty} w_k(s) = \begin{cases} 0\,, & 0 \le s < 1\,, \\ \|v\|\,, & s = 1\,, \end{cases}$$

for any $v \in \mathcal{N}(A)^\perp$.

Indeed, in the linear and noiseless case ($F(x) = Kx$, $\delta = 0$) we get from (45) using $Kx^\dagger = y$ and (8)

$$\begin{aligned} x_{k+1} - x^\dagger &= x_k - x^\dagger + (K^*K + \alpha_k I)^{-1}(K^*K(x^\dagger - x_k) + \alpha_k(x_0 - x^\dagger + x^\dagger - x_k)) \\ &= -\alpha_k(K^*K + \alpha_k I)^{-1}(K^*K)^\mu v \end{aligned}$$

To take into account noisy data and nonlinearity, we rewrite (45) as

$$\begin{aligned} x_{k+1}^\delta - x^\dagger = &-\alpha_k(K^*K + \alpha_k I)^{-1}(K^*K)^\mu v \\ &-\alpha_k(K_k^*K_k + \alpha_k I)^{-1}\left(K^*K - K_k^*K_k\right) \\ &\quad (K^*K + \alpha_k I)^{-1}(K^*K)^\mu v \\ &+ (K_k^*K_k + \alpha_k I)^{-1}K_k^*(y^\delta - F(x_k^\delta) + K_k(x_k^\delta - x^\dagger))\,, \end{aligned} \tag{48}$$

where we set $K_k := F'(x_k^\delta)$, $K := F'(x^\dagger)$.

Let us consider the case that $0 \le \mu < 1/2$ in (8) and assume that the nonlinearity condition (44) as well as $x_k^\delta \in \mathcal{B}_\rho(x^\dagger) \subseteq \mathcal{B}_{2\rho}(x_0)$ hold. Therewith, for the Taylor remainder we obtain that

$$\|F(x_k^\delta) - F(x^\dagger) - K_k(x_k^\delta - x^\dagger)\| \le 2c_R \|K(x_k^\delta - x^\dagger)\| . \tag{49}$$

The estimates [see (47)]

$$\|(K_k^* K_k + \alpha_k I)^{-1}\| \le \alpha_k^{-1} , \qquad \|(K_k^* K_k + \alpha_k I)^{-1} K_k^*\| \le \tfrac{1}{2}\alpha_k^{-\frac{1}{2}} ,$$

and the identity

$$K^*K - K_k^*K_k = K_k^*(R_{x_k^\delta}^{-1^*} - R_{x_k^\delta})K$$

imply that

$$\begin{aligned}&\|\alpha_k(K_k^*K_k + \alpha_k I)^{-1}(K^*K - K_k^*K_k)(K^*K + \alpha_k I)^{-1}(K^*K)^\mu v\| \\ &\le \tfrac{1}{2}\alpha_k^{-\frac{1}{2}} \|R_{x_k^\delta}^{-1^*} - R_{x_k^\delta}\| \, \|K(K^*K + \alpha_k I)^{-1}(K^*K)^\mu v\|\end{aligned}$$

This together with (2), (47), (48), and $F(x^\dagger) = y$ yields the estimate (50) in Lemma 2 below. Inserting the identity $K = R_{x_k^\delta}^{-1} K_k$ into (48) we obtain

$$\begin{aligned}Ke_{k+1}^\delta = &-\alpha_k K(K^*K + \alpha_k I)^{-1}(K^*K)^\mu v \\ &-\alpha_k R_{x_k^\delta}^{-1} K_k(K_k^*K_k + \alpha_k I)^{-1} K_k^*(R_{x_k^\delta}^{-1^*} - R_{x_k^\delta})K \\ &\quad (K^*K + \alpha_k I)^{-1}(K^*K)^\mu v \\ &- R_{x_k^\delta}^{-1} K_k(K_k^*K_k + \alpha_k I)^{-1} K_k^* \\ &\quad (F(x_k^\delta) - F(x^\dagger) - K_k(x_k^\delta - x^\dagger) + y - y^\delta) .\end{aligned}$$

Now the estimate (51) in Lemma 2 below follows together with (2), (44), (47), and (49).

Similarly one can derive estimates (52), (53) in case of $1/2 \le \mu \le 1$ under the Lipschitz condition (4), by using (5) and the decomposition

$$K^*K - K_k^*K_k = K_k^*(K - K_k) + (K^* - K_k^*)K .$$

Lemma 2. *Let* (3), (8), (46) *hold and assume that* $x_k^\delta \in \mathcal{B}_\rho(x^\dagger)$. *Moreover, set* $K := F'(x^\dagger)$, $e_k^\delta := x_k^\delta - x^\dagger$, *and let* $w_k(\cdot)$ *be defined as in* (47).

(i) If $0 \le \mu < 1/2$ *and* (44) *hold, we obtain the estimates*

$$\|e_{k+1}^\delta\| \le \alpha_k^\mu w_k(\mu) + c_R \alpha_k^\mu w_k(\mu+\tfrac{1}{2}) + \alpha_k^{-\frac{1}{2}}(c_R\|Ke_k^\delta\| + \tfrac{1}{2}\delta)\,, \quad (50)$$

$$\begin{aligned}\|Ke_{k+1}^\delta\| &\le (1+2c_R(1+c_R))\alpha_k^{\mu+\frac{1}{2}} w_k(\mu+\tfrac{1}{2}) \\ &\quad + (1+c_R)\,(2c_R\|Ke_k^\delta\| + \delta)\,. \end{aligned} \quad (51)$$

(ii) If $1/2 \le \mu \le 1$ *and* (4) *hold, we obtain the estimates*

$$\begin{aligned}\|e_{k+1}^\delta\| &\le \alpha_k^\mu w_k(\mu) + L\|e_k^\delta\|(\tfrac{1}{2}\alpha_k^{\mu-\frac{1}{2}} w_k(\mu) + \|(K^*K)^{\mu-\frac{1}{2}}v\|) \\ &\quad + \tfrac{1}{2}\alpha_k^{-\frac{1}{2}}(\tfrac{1}{2}L\|e_k^\delta\|^2 + \delta)\,, \end{aligned} \quad (52)$$

$$\begin{aligned}\|Ke_{k+1}^\delta\| &\le \alpha_k\|(K^*K)^{\mu-\frac{1}{2}}v\| + L^2\|e_k^\delta\|^2(\tfrac{1}{2}\alpha_k^{\mu-\frac{1}{2}} w_k(\mu) + \|(K^*K)^{\mu-\frac{1}{2}}v\|) \\ &\quad + L\alpha_k^{\frac{1}{2}}\|e_k^\delta\|(\alpha_k^{\mu-\frac{1}{2}} w_k(\mu) + \tfrac{1}{2}\|(K^*K)^{\mu-\frac{1}{2}}v\|) \\ &\quad + (\tfrac{1}{2}L\alpha_k^{-\frac{1}{2}}\|e_k^\delta\| + 1)(\tfrac{1}{2}L\|e_k^\delta\|^2 + \delta)\,. \end{aligned} \quad (53)$$

It is readily checked that the nonlinearity condition (44) used in Lemma 2 can be extended to

$$F'(\tilde{x}) = R(\tilde{x},x)F'(x) + Q(\tilde{x},x) \quad (54)$$

$$\|I - R(\tilde{x},x)\| \le c_R \quad (55)$$

$$\|Q(\tilde{x},x)\| \le c_Q\|F'(x^\dagger)(\tilde{x}-x)\| \quad (56)$$

for $x, \tilde{x} \in \mathcal{B}_{2\rho}(x_0)$, where c_R and c_Q are nonnegative constants.

With the a-priori stopping rule

$$k_* \to \infty \qquad \text{and} \qquad \eta \ge \delta\alpha_{k_*}^{-\frac{1}{2}} \to 0 \ \text{ as } \ \delta \to 0\,. \quad (57)$$

for $\mu = 0$ and

$$\eta\alpha_{k_*}^{\mu+\frac{1}{2}} \le \delta < \eta\alpha_k^{\mu+\frac{1}{2}}\,, \qquad 0 \le k < k_*\,, \quad (58)$$

for $0 < \mu \le 1$ one obtains optimal convergence rates as follows (see [70, Theorem 4.12]):

Theorem 9. *Let* (3), (8), (46) *hold and let* $k_* = k_*(\delta)$ *be chosen according to* (57) *for* $\mu = 0$ *and* (58) *for* $0 < \mu \le 1$, *respectively.*

(i) If $0 \le \mu < 1/2$, *we assume that* (54)–(56) *hold and that* $\|x_0 - x^\dagger\|$, $\|v\|$, η, ρ, c_R *are sufficiently small.*

(ii) If $1/2 \le \mu \le 1$, *we assume that* (4) *and* $\|x_0 - x^\dagger\|$, $\|v\|$, η, ρ *are sufficiently small.*

Then we obtain that

$$\|x_{k_*}^\delta - x^\dagger\| = \begin{cases} o(1)\,, & \mu = 0\,, \\ O\left(\delta^{\frac{2\mu}{2\mu+1}}\right), & 0 < \mu \le 1\,. \end{cases}$$

For the noise free case ($\delta = 0$, $\eta = 0$) *we obtain that*

$$\|x_k - x^\dagger\| = \begin{cases} o(\alpha_k^\mu)\,, & 0 \le \mu < 1\,, \\ O(\alpha_k)\,, & \mu = 1\,, \end{cases}$$

and that

$$\|F(x_k) - y\| = \begin{cases} o\left(\alpha_k^{\mu+\frac{1}{2}}\right), & 0 \le \mu < \frac{1}{2}\,, \\ O(\alpha_k)\,, & \frac{1}{2} \le \mu \le 1\,. \end{cases}$$

With the discrepancy principle (10) as an a-posteriori stopping rule in place of the a-priori stopping rule (57), (58), optimal rates can be obtained under a Hölder type source condition (8) with $\mu \le \frac{1}{2}$ (see [70, Theorem 4.13]):

Theorem 10. *Let* (3), (8), (46), *and* (54)–(56) *hold for some* $0 \le \mu \le 1/2$, *and let* $k_* = k_*(\delta)$ *be chosen according to* (10) *with* $\tau > 1$. *Moreover, we assume that* $\|x_0 - x^\dagger\|$, $\|v\|$, $1/\tau$, ρ, *and* c_R *are sufficiently small. Then we obtain the rates*

$$\|x_{k_*}^\delta - x^\dagger\| = \begin{cases} o\left(\delta^{\frac{2\mu}{2\mu+1}}\right), & 0 \le \mu < \frac{1}{2}\,, \\ O(\sqrt{\delta})\,, & \mu = \frac{1}{2}\,. \end{cases}$$

In case $\mu = 0$, and with an a-posteriori choice of α_k similar to (35), the nonlinearity condition can be relaxed to (6), see [71].

Further Literature on Gauss–Newton Type Methods

Generalizations of the IRGNM

Already Bakushinskii in [4] proposed to replace Tikonov regularization in (45) by a more general method defined via functional calculus by a filter function g with $g(\lambda) \approx \frac{1}{\lambda}$:

$$x_{k+1}^\delta = x_0 + g(F'(x_k^\delta)^* F'(x_k^\delta)) F'(x_k^\delta)^* (y^\delta - F(x_k^\delta) - F'(x_k^\delta)(x_0 - x_k^\delta))\,, \qquad (59)$$

with $\alpha_k \searrow 0$.

Still more general, one can replace the operator $g(F'(x_k^\delta)^* F'(x_k^\delta))F'(x_k^\delta)^*$ by some regularization operator $R_\alpha(F'(x))$ with

$$R_\alpha(F'(x)) \approx F'(x)^\dagger ,$$

satisfying certain structural conditions so that the convergence analysis for the resulting Newton type method

$$x_{k+1}^\delta = x_0 + R_{\alpha_k}(F'(x_k^\delta))(y^\delta - F(x_k^\delta) - F'(x_k^\delta)(x_0 - x_k^\delta)) . \tag{60}$$

(see [30, 55, 59, 61, 62, 70]) applies not only to methods defined via functional calculus such as iterated Tikhonov regularization, Landweber iteration, and Lardy's method but also to regularization by discretization.

An additional augmentation of the analysis concerns the type of nonlinearity condition. Alternatively to range invariance of the adjoint of $F'(x)$ (41), which is closely related to (6), one can consider range invariance of $F'(x)$ itself

$$F'(\tilde{x}) = F'(x)R(\tilde{x}, x) \qquad \text{and} \quad \|I - R(\tilde{x}, x)\| \le c_R \|\tilde{x} - x\| \tag{61}$$

for $x, \tilde{x} \in \mathcal{B}_{2\rho}(x_0)$ and some positive constant c_R.

Incorporation of convex constraints is considered in [66, 102].

Generalized Source Conditions

Convergence and optimal rates for the iteratively regularized Gauss–Newton method were established in [82] under a general source condition of the form

$$x^\dagger - x_0 = f(F'(x^\dagger)^* F'(x^\dagger))v , \quad v \in \mathcal{N}(F'(x^\dagger))^\perp ,$$

with an index function $f : [0, \|F'(x^\dagger)\|^2] \to [0, \infty]$ that is increasing and continuous with $f(0) = 0$. These include logarithmic source conditions (9) that are appropriate for severely ill-posed problems. For this purpose, it is assumed that conditions (54)–(56) hold and the iteration is stopped according to the discrepancy principle (10).

Other A-posteriori Stopping Rules

Bauer and Hohage in [6] carry out a convergence analysis with the Lepskii balancing principle, i.e.,

$$k_* = \min \Big\{ k \in \{0, \dots, k_{\max}\} : \|x_k^\delta - x_m^\delta\| \le 8 c_\Phi \alpha_m^{-1/2} \delta \\ \forall m \in \{k+1, \dots, k_{\max}\} \Big\} \tag{62}$$

(with $k_{\max} = k_{\max}(\delta)$ an a-priori determined index up to which the iterates are well defined) in place of the discrepancy principle as an a-posteriori stopping rule.

Optimal convergence rates are shown for (60) with R_α defined by Landweber iteration or (iterated) Tikhonov regularization under a condition similar to (54)–(56) if (8) with $\mu \leq 1/2$ or (9) holds, and under condition (4) if $\mu \geq 1/2$ in (8). The advantage of this stopping rule is that saturation at $\mu = 1/2$ is avoided.

Stochastic Noise Models

In many practical applications (e.g., in weather forecast), the data noise is not only of deterministic nature as assumed in our exposition, but also random noise has to be taken into account. In [8] Bauer, Hohage, and Munk consider the noise model

$$y^{\delta,\sigma} = F(x^\dagger) + \delta\eta + \sigma\xi$$

where $\eta \in \mathcal{Y}$, $\|\eta\| \leq 1$ describes the deterministic part of the noise with noise level δ, ξ is a normalized Hilbert space process in $\mathcal{Y}$ (see, e.g., [13]) and σ^2 is the variance of the stochastic noise. Under a Hölder source condition (8) with $\mu > 1/2$ and assuming a Lipschitz condition (4), they show almost optimal convergence rates (i.e., with an additional factor that is logarithmic in σ) of (60) with R_α defined by iterated Tikhonov regularization and with the Lepskii balancing principle (62) as a stopping rule. The setting of Hohage and Werner [57] allows for an even much more general setting with regard to the stochastic noise.

Generalization to Banach Space

Bakushinski and Kokurin in [5] consider the setting $\mathcal{Y} = \mathcal{X}$ with $\mathcal{X}$ Banach space. Using the Riesz–Dunford formula, they prove optimal convergence rates for the generalized Newton method (59) under the Lipschitz condition (4), provided a sufficiently strong source condition, namely (8) with $\mu \geq 1/2$ holds.

In [71], based on the variational formulation of the iteratively regularized Gauss–Newton method,

$$x_{k+1}^\delta = \arg\min_{x \in \mathcal{D}(F)} \|y^\delta - F(x_k^\delta) - F'(x_k^\delta)(x - x_k^\delta)\|^p + \alpha_k \|x - x_0\|^q$$

with $p, q \in (1, \infty)$ and Banach space norms, convergence in the general situation of possibly different Banach spaces $\mathcal{X}$, $\mathcal{Y}$ without source condition under the nonlinearity assumption (6) is proved. Convergence rates under variational and approximate source conditions generalizing (8) to the Banach space setting are provided in the paper [65].

The convergence rates results in [57] even hold for more general data misfit as well as regularization functionals in place of Banach space norms. Results on Gauss–Newton methods with other regularization than Tikhonov (similarly to section "Generalizations of the IRGNM") can be found, e.g., in [58, 68].

Efficient Implementation

To speed up convergence and save computational effort, it is essential to use preconditioning when applying an iterative regularization method R_α in (60).

Egger in [32] defines preconditioners for these iterations (Landweber iteration, CG, or the ν-methods, see, e.g., [36]) via Hilbert scales (see, e.g., [36]), which leads to an iterative scheme of the form

$$x_{k+1}^{\delta} = x_0 + g(\mathcal{L}^{-2s} F'(x_k^{\delta})^* F'(x_k^{\delta})) \mathcal{L}^{-2s} F'(x_k^{\delta})^* (y^{\delta} - F(x_k^{\delta}) - F'(x_k^{\delta})(x_0 - x_k^{\delta})) ,$$

where $\mathcal{L}$ is typically a differential operator and s an appropriately chosen exponent. It is shown in [32] that this leads to a reduction of the number of iterations to about the square root.

In his thesis [81], Langer makes use of the close connection between the CG iteration and Lanczos' method in order to construct a spectral preconditioner that is especially effective for severely ill-posed problems.

Further strategies for saving computational effort are, e.g., multigrid [1, 42, 63, 67, 76], quasi Newton [41, 60], and adaptive discretization [72, 73] methods.

5 Nonstandard Iterative Methods

The methods presented above were based on the standard ideas of minimizing a least-squares functional, namely gradient descent and Newton methods. In the following we shall discuss further iterative methods, either not based on descent of the objective functional or based on descent for a different functional than least-squares.

Kaczmarz and Splitting Methods

Kaczmarz-type methods are used as splitting algorithms for large operators. They are usually applied if $\mathcal{Y}$ and F can be split into

$$\mathcal{Y} = \mathcal{Y}_1 \times \mathcal{Y}_2 \times \cdots \times \mathcal{Y}_M$$

and

$$F = (F_1, F_2, \ldots, F_M) ,$$

with continuous operators $F_j : \mathcal{X} \to \mathcal{Y}_j$. The corresponding least-squares problem is the minimization of the functional

$$J(x) = \tfrac{1}{2} \sum_{j=1}^{M} \| F_j(x) - y_j^{\delta} \|_{\mathcal{Y}_j}^2 .$$

The basic idea of a Kaczmarz-type method is to apply an iterative scheme to each of the least-squares terms $\frac{1}{2} \| F_j(x) - y_j^{\delta} \|_{\mathcal{Y}_j}^2$ separately in substeps of the iteration. The three most commonly used approaches are the *Landweber–Kaczmarz method* (cf. [77])

$$x^\delta_{k+j/M} = x^\delta_{k+(j-1)/M} - \omega_k F_j'(x^\delta_{k+(j-1)/M})^*(F_j(x^\delta_{k+(j-1)/M}) - y^\delta_j)\,,$$
$$j = 1, \dots, M$$

the *nonlinear Kaczmarz method*

$$x^\delta_{k+j/M} = x^\delta_{k+(j-1)/M} - \omega_k F_j'(x^\delta_{k+j/M})^*(F_j(x^\delta_{k+j/M}) - y^\delta_j)\,, \qquad j = 1, \dots, M$$

and the *Gauss–Newton–Kaczmarz method*

$$x^\delta_{k+j/M} = x^\delta_{k+(j-1)/M} - (F_j'(x^\delta_{k+(j-1)/M})^* F_j'(x^\delta_{k+(j-1)/M}) + \alpha_{k,j} I)^{-1}$$
$$F_j'(x^\delta_{k+(j-1)/M})^*(F_j(x^\delta_{k+(j-1)/M}) - y^\delta_j)\,.$$

Further Newton–Kaczmarz methods can be constructed in the same way as iteratively regularized and inexact Newton methods (cf. [19]).

The Landweber–Kaczmarz and the nonlinear Kaczmarz method can be interpreted as time discretization by operator splitting for the minimizing flow

$$x'(t) = -\sum_{j=1}^{M} F_j'(x(t))^*(F_j(x(t) - y^\delta_j)\,,$$

with forward, respectively, backward Euler operator splitting (cf. [38, 92]). The nonlinear Kaczmarz method is actually a special case of the *Douglas–Rachford splitting algorithm* applied to the above least-squares problem, the iterate $x^\delta_{k+j/M}$ can be computed as a minimizer of the Tikhonov-type functional

$$J_{k,j}(x) = \tfrac{1}{2}\|F_j(x) - y^\delta_j\|^2_{\mathcal{Y}_j} + \tfrac{1}{2\tau}\|x - x^\delta_{k+(j-1)/M}\|^2.$$

The convergence analysis of Kaczmarz methods is very similar to the analysis of the iterative methods mentioned above, if nonlinearity conditions on each single operator F_j are posed (cf. [43, 44, 77] for the Landweber–Kaczmarz, [10, 19, 83] for Newton–Kaczmarz, [9, 27] for nonlinear Kaczmarz and further variants). The verification of those conditions is usually an even harder task than for the collection of operators $F = (F_1, \dots, F_M)$, also due to the usually large nullspace of their linearizations. The analysis can however provide at least a good idea on the convergence behavior of the algorithms. A nontrivial point in Kaczmarz methods is an a-posteriori stopping criterion, since in general the overall residual is not decreasing, which rules out standard approaches such as the discrepancy principles. Some discussions of this issue can be found in [43], where criteria based on the sequence of residuals $(\|F(x^\delta_{k+j/M} - y^\delta_j\|)_{j=1,\dots,M}$ have been introduced, supplemented by additional skipping strategies.

Kaczmarz methods have particular advantages in inverse problems for partial differential equations, when many state equations for different parameters (e.g., different boundary values or different sources) need to be solved. Then the operators

F_j can be set up such that a problem for one state equation can be solved after the other, which is memory efficient. We mention that in this case also the Landweber iteration can be carried out in the same memory-efficient way, since

$$F'(x)^*(F(x) - y^\delta) = \sum_{j=1}^{M} F_j'(x)^*(F_j(x) - y_j^\delta)\,.$$

But in most cases one observes faster convergence for the Kaczmarz-type variant, which is similar as comparing classical Jacobi and Gauss–Seidel methods.

Splitting methods are frequently used for the iterative solution of problems with variational regularization of the form

$$x_\alpha^\delta \in \arg\min_x \left[\tfrac{1}{2}\|F(x) - y^\delta\|^2 + \alpha R(x)\right],$$

where $R : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ is an appropriate convex regularization functional. It is then natural to apply operator splitting to the least-squares part and the regularization part, if R is not quadratic. The most important approaches are the *Douglas–Rachford splitting* (in particular for linear operators F, cf. [31])

$$x_{k+1/2}^\delta \in \arg\min_x \left[\tfrac{1}{2}\|F(x) - y\|^2 + \tfrac{1}{2\omega_k}\|x - x_k^\delta\|^2\right]$$
$$x_{k+1}^\delta \in \arg\min_x \left[R(x) + \tfrac{1}{2\omega_k}\|x - x_{k+1/2}^\delta\|^2\right]$$

and the *forward–backward splitting algorithm* (cf. [84])

$$x_{k+1/2}^\delta = x_k^\delta - \omega_k F'(x_k^\delta)^*(F(x_k^\delta) - y^\delta)$$
$$x_{k+1}^\delta \in \arg\min_x \left[R(x) + \frac{1}{2\omega_k}\|x - x_{k+1/2}^\delta\|^2\right]$$

Such algorithms are particularly popular for nonsmooth regularization (cf. [25]), in the case of sparsity enforcing penalties (ℓ^1-norms) the second step in both algorithms can be computed explicitly via shrinkage (thresholding) formulas, such schemes are hence also called *iterative shrinkage* (cf. [26]).

EM Algorithms

A very popular algorithm in the case of image reconstruction with nonnegativity constraints is the *expectation maximization (EM) method*, also called *Richardson–Lucy algorithm* (cf. [12, 86]). In the case of $F : L^1(\Omega) \to L^1(\Sigma)$ being a linear operator, it is given by the multiplicative fixed-point scheme

$$x_{k+1}^\delta = x_k^\delta F^*\left(\frac{y^\delta}{F x_k^\delta}\right). \tag{63}$$

For F and F^* being positivity preserving operators (such as the Radon transform or convolutions with positive kernels), the algorithm preserves the positivity of an initial value x_0^δ if the data y^δ are positive, too. Positivity of data is a too strong restriction in the case of an additive noise model, like stochastic Gaussian models or bounds in squared L^2-distances. It is however well suited for multiplicative models such as Poisson models used for imaging techniques based on counting emitted particles (photons or positrons). The log-likelihood functional of a Poisson model, respectively its asymptotic for large count rates, can also be used to derive a variational interpretation of the EM-algorithm. More precisely (63) is a descent method for the functional

$$J(x) := \int_\Sigma \left[y^\delta \log\left(\frac{y^\delta}{Fx}\right) - y^\delta + Fx \right] d\sigma \,,$$

which corresponds to the *Kullback–Leibler divergence* (relative entropy) between the output Fx and the data y^δ. Minimizing J over nonnegative functions leads to the optimality condition

$$x\left(-F^*\left(\frac{y^\delta}{Fx}\right) + F^*1\right) = 0\,.$$

With appropriate operator scaling $F^*1 = 1$ this yields the fixed-point equation

$$x = xF^*\left(\frac{y^\delta}{Fx}\right),$$

which is the basis of the EM-algorithm:

$$x_{k+1}^\delta = \frac{x_k^\delta}{F^*1} F^*\left(\frac{y^\delta}{Fx_k^\delta}\right).$$

The EM-algorithm or Richardson–Lucy algorithm (cf. [103]) is a special case of the general EM framework by Dempster, Laird, and Rubin (cf. [29]).

Performing an analogous analysis for a nonlinear operator $F : L^1(\Omega) \to L^1(\Sigma)$ we are led to the fixed-point equation

$$xF'(x)^*1 = xF'(x)^*\left(\frac{y^\delta}{Fx}\right).$$

Since it seems unrealistic to scale $F'(x)^*$ for arbitrary x it is more suitable to keep the term and divide by $F'(x)^*1$. The corresponding fixed-point iteration is the *nonlinear EM algorithm*

$$x_{k+1}^\delta = \frac{x_k^\delta}{F'(x_k^\delta)^*1} F'(x_k^\delta)^*\left(\frac{y^\delta}{F(x_k^\delta)}\right).$$

The convergence analysis in the nonlinear case is still widely open. Therefore, we only comment on the case of linear operators F here (cf. also [85,86] for details). In order to avoid technical difficulties the scaling condition $F^*1 = 1$ will be assumed in the following. The major ingredients are reminiscent of the convergence analysis of the Landweber iteration, but they replace the norm distance by the Kullback–Leibler divergence

$$KL(x, \tilde{x}) = \int_\Omega \left[x \log \frac{x}{\tilde{x}} - x + \tilde{x} \right] d\omega,$$

which is a nonnegative distance, but obviously not a metric. First of all, x_k^δ can be characterized as a minimizer of the (convex) functional

$$J_k(x) := -J(x) + KL(x_{k+1}^\delta, x)$$

over all nonnegative L^1-functions, given x_{k+1}^δ. Thus, comparing with the functional value at x_{k+1}^δ we find that the likelihood functional J decreases during the iteration, more precisely

$$J(x_{k+1}^\delta) \leq J(x_k^\delta) - KL(x_{k+1}^\delta, x_k^\delta)\,.$$

This part of the analysis holds also in the case of exact data. The second inequality directly concerns the dissipation of the Kullback–Leibler divergence between the iterates and a solution, hence assumes the existence of $x^\dagger$ with $Fx^\dagger = y$. Using convexity arguments one obtains

$$KL(x^\dagger, x_{k+1}) + J(x_k) \leq KL(x^\dagger, x_k)\,.$$

Hence, together with the monotonicity of $(J(x_j))_{j\in\mathbb{N}}$

$$KL(x^\dagger, x_k) + kJ(x_k) \leq KL(x^\dagger, x_k) + \sum_{j=0}^{k-1} J(x_j) \leq KL(x^\dagger, x_0)\,,$$

which implies boundedness of the Kullback–Leibler divergence (hence a weak compactness of x_k in L^1) and convergence $J(x_k) \to 0$ analogous to the arguments for the Landweber iteration.

The noisy case is less clear, apparently also due to the difficulties in defining a reasonable noise level for Poisson noise. An analysis defining the noise level in terms of the likelihood of the noisy data has been given in [93]. Further analysis in the case of noisy data seems to be necessary, however. This also concerns stopping rules for noisy data, which are usually based on the noise level. A promising multiscale stopping criterion based on the stochastic modeling of Poisson noise has been introduced and tested recently (cf. [14]). For a combination of EM with Kaczmarz ideas, see [45].

Iterative methods are also used for Penalized EM-Reconstruction (equivalently Bayesian MAP estimation), i.e., for minimizing

$$J(x) + \alpha R(x)$$

over nonnegative L^1-functions, where $\alpha > 0$ is a regularization parameter and R is an appropriate regularization functional, e.g., total variation or negative entropy.

A frequently used modification of the EM algorithm in this case is Green's One-Step-Late (OSL) method (see [39,40]).

$$x_{k+1}^{\delta} = \frac{x_k^{\delta}}{F^*1 + \alpha R'(x_k)} F^* \left(\frac{y^{\delta}}{F x_k^{\delta}} \right),$$

which seems efficient if the pointwise sign of $R'(x_k)$ can be controlled, e.g., for entropy-type regularization functionals

$$R(x) = \int_{\Omega} E(x)$$

with convex $E : \Omega \to \mathbb{R}^+$ and $E'(0) = 0$. The additional effort compared to EM is negligible and the method converges reasonably fast to a minimizer of $J + \alpha R$. For other important variational regularization methods, in particular gradient-based functionals, the OSL method is less successful, since $R'(x_k)$ does not necessarily have the same sign as x_k, thus $F^*1 + \alpha R'(x_k)$ can be negative or zero, in which case the iteration has to be stopped or some ad-hoc fixes have to be introduced. Another obvious disadvantage of the OSL method is the fact that it cannot handle nonsmooth regularizations such as total variation and ℓ^1-norms, which are often used to incorporate structural prior knowledge. As a more robust alternative, splitting methods have been introduced also in this case. In [97] a positivity-preserving forward–backward splitting algorithm with particular focus on total variation regularization has been introduced. The two-step algorithm alternates the classical EM-step with a weighted denoising problem

$$x_{k+1/2}^{\delta} = x_k^{\delta} F^* \left(\frac{y^{\delta}}{F x_k^{\delta}} \right)$$

$$x_{k+1}^{\delta} \in \arg\min_x \left[\int_{\Omega} \frac{(x - x_{k+1/2}^{\delta})^2}{x_k^{\delta}} + \alpha R(x) \right].$$

Convergence can be ensured with further damping, i.e., if the second half step is replaced by

$$x_{k+1}^{\delta} \in \arg\min_x \left[\int_{\Omega} \frac{(x - \omega_k x_{k+1/2}^{\delta} - (1 - \omega_k) x_k^{\delta})^2}{x_k^{\delta}} + \alpha R(x) \right]$$

with $\omega_k \in (0,1)$ sufficiently small. This algorithm is a semi-implicit approximation of the optimality condition

$$-F^*\left(\frac{y^\delta}{Fx}\right) + 1 + \alpha p = 0, \quad p \in \partial R(x)\,,$$

where the operator-dependent first part is approximated explicitly and the regularization part p implicitly. What seems surprising is that the constant 1 is approximated by $x^\delta_{k+1}/x^\delta_k$, which however turns out to be crucial for preserving positivity.

Bregman Iterations

A very general way of constructing iterative methods in Banach spaces are iterations using so-called *Bregman distances*. For a convex functional R, the Bregman distance is defined by

$$D_R^p(\tilde{x}, x) = R(\tilde{x}) - R(x) - \langle\, p, \tilde{x} - x \,\rangle\,, \quad p \in \partial R(x)\,.$$

Note that for nonsmooth R the subgradient is not single-valued, hence the distance depends on the choice of the specific subgradient. Bregman distances are a very general class of distances in general, the main properties are $D_R^p(\tilde{x}, x) \geq 0$ and $D_R^p(x, x) = 0$. Particular cases are

$$D_R^p(\tilde{x}, x) = \tfrac{1}{2}\,\|\tilde{x} - x\|^2 \quad \text{for} \quad R(x) = \tfrac{1}{2}\,\|x\|^2$$

and the Kullback–Leibler divergence for R being a logarithmic entropy functional.

If some data similarity measure $H(F(x), y^\delta)$ and a regularization functional R is given, the Bregman iteration (cf. [16,91] in its original, different context) consists of

$$x^\delta_{k+1} \in \arg\min_x \left[H(F(x), y^\delta) + D_R^{p_k}(x, x^\delta_k)\right]$$

with the dual update

$$p_{k+1} = p_k - \partial_x H(F(x^\delta_{k+1}), y^\delta) \in \partial R(x^\delta_{k+1})\,.$$

The Bregman iteration is a primal–dual method in the sense that it computes an update for the primal variable x as well as for the dual variable $p \in \partial R(x)$. Consequently one also needs to specify an initial value for the subgradient $p_0 \in \partial R(x^\delta_0)$.

Most investigations of the Bregman iteration have been carried out for H being a squared norm, i.e., the least-squares case discussed above

$$H(F(x), y^\delta) = \tfrac{1}{2}\,\|F(x) - y^\delta\|^2.$$

Under appropriate nonlinearity conditions a full convergence analysis can be carried out (cf. [2]), in general only leading to some weak convergence and convergence in the Bregman distance. If F is a nonlinear operator, further approximations in the Bregman iterations by linearization are possible leading to the Landweber-type method (also called linearized Bregman iteration)

$$x_{k+1}^{\delta} \in \arg\min_{x} \left[\langle F'(x_k^{\delta})(x - x_k^{\delta}), F(x_k^{\delta}) - y^{\delta} \rangle + D_R^{p_k}(x, x_k^{\delta}) \right]$$

and Levenberg–Marquardt type method

$$x_{k+1}^{\delta} \in \arg\min_{x} \left[\tfrac{1}{2} \| F(x_k^{\delta}) + F'(x_k^{\delta})(x - x_k^{\delta}) - y^{\delta} \|^2 + D_R^{p_k}(x, x_k^{\delta}) \right].$$

Both schemes have been analyzed in [2], see also [21–23] for the linearized Bregman iteration in compressed sensing. We mention that in particular the linearized Bregman method does not work with an arbitrary convex regularization functional. In order to guarantee that the functional to be minimized in each step of the iteration is bounded from below so that the iterates are well defined, a quadratic part in the regularization term is needed.

A discussion of Bregman iterations in the case of nonquadratic term H can be found in [17, 18, 50] with particular focus on F being a linear operator. In this case also a dual Bregman iteration can be constructed, which coincides with the original one in the case of quadratic H, but differs in general. For this dual Bregman iterations also convergence rates under appropriate source conditions can be shown (cf. [18]), which seems out of reach for the original Bregman iteration for general H. A systematic analysis of Bregman iterations in image restoration can be found in [20]. In [11], Bregman iterations are used to enhance generalized total variation and infimal convolution regularization.

6 Conclusion

Iterative methods offer an attractive alternative to variational regularization but are also closely linked to them via iterative optimization. In this chapter we aimed at giving a broad overview on the main classical (gradient and Newton type) as well as nonstandard (Kaczmarz, expectation maximization, Bregman) iterations. We put an emphasis on their regularizing properties for nonlinear ill-posed problems in Hilbert spaces and provided outlooks on further aspects such as efficient implementation, stochastic noise models, or formulations in Banach spaces.

Cross-References

▸ EM Algorithms
▸ EM Algorithms from a Non-stochastic Perspective

- Linear Inverse Problems
- Regularization Methods for Ill-Posed Problems

References

1. Akcelik, V., Biros, G., Draganescu, A., Hill, J., Ghattas, O., Waanders, B.V.B.: Dynamic data-driven inversion for terascale simulations: real-time identification of airborne contaminants. In: Proceedings of SC05. IEEE/ACM, Seattle (2005)
2. Bachmayr, M., Burger, M.: Iterative total variation schemes for nonlinear inverse problems. Inverse Prob. **25**, 105,004 (2009)
3. Bakushinsky, A.B.: The problem of the convergence of the iteratively regularized Gauss-Newton method. Comput. Math. Math. Phys. **32**, 1353–1359 (1992)
4. Bakushinsky, A.B.: Iterative methods without degeneration for solving degenerate nonlinear operator equations. Dokl. Akad. Nauk. **344**, 7–8 (1995)
5. Bakushinsky, A.B., Kokurin, M.Y.: Iterative Methods for Approximate Solution of Inverse Problems. Mathematics and Its Applications, vol. 577. Springer, Dordrecht (2004)
6. Bauer, F., Hohage, T.: A Lepskij-type stopping rule for regularized Newton methods. Inverse Prob. **21**, 1975–1991 (2005)
7. Bauer, F., Kindermann, S.: The quasi-optimality criterion for classical inverse problems. Inverse Prob. **24**(3), 035,002 (20 pp) (2008)
8. Bauer, F., Hohage, T., Munk, A.: Iteratively regularized Gauss-Newton method for nonlinear inverse problems with random noise. SIAM J. Numer. Anal. **47**, 1827–1846 (2009)
9. Baumeister, J., Cezaro, A.D., Leitao, A.: On iterated Tikhonov-Kaczmarz regularization methods for ill-posed problems. ICJV (2010). doi:10.1007/s11263-010-0339-5
10. Baumeister, J., Kaltenbacher, B., Leitao, A.: On Levenberg-Marquardt Kaczmarz regularization methods for ill-posed problems. Inverse Prob. Imaging **4**, 335–350 (2010)
11. Benning, M., Brune, C., Burger, M., Mueller, J.: Higher-order tv methods - enhancement via bregman iteration. J. Sci. Comput. **54**(2-3), 269–310 (2013). In honor of Stanley Osher for his 70th birthday
12. Bertero, M., Boccacci, P.: Introduction to Inverse Problems in Imaging. Institute of Physics Publishing, Bristol (1998)
13. Bissantz, N., Hohage, T., Munk, A., Ruymgaart, F.: Convergence rates of general regularization methods for statistical inverse problems and applications. SIAM J. Numer. Anal. **45**, 2610–2636 (2007)
14. Bissantz, N., Mair, B., Munk, A.: A statistical stopping rule for MLEM reconstructions in PET. IEEE Nucl. Sci. Symp. Conf. Rec **8**, 4198–4200 (2008)
15. Blaschke, B., Neubauer, A., Scherzer, O.: On convergence rates for the iteratively regularized Gauss-Newton method. IMA J. Numer. Anal. **17**, 421–436 (1997)
16. Bregman, L.M.: The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. USSR Comp. Math. Math. Phys. **7**, 200–217 (1967)
17. Brune, C., Sawatzky, A., Burger, M.: Bregman-EM-TV methods with application to optical nanoscopy. In: Tai, X.-C., et al. (ed.) Proceedings of the 2nd International Conference on Scale Space and Variational Methods in Computer Vision. Lecturer Notes in Computer Science, vol. 5567, pp. 235–246. Springer, New York (2009)
18. Brune, C., Sawatzky, A., Burger, M.: Primal and dual Bregman methods with application to optical nanoscopy. Int. J. Comput. Vis **92**, 211–229 (2011)
19. Burger, M., Kaltenbacher, B.: Regularizing Newton-Kaczmarz methods for nonlinear ill-posed problems. SIAM J. Numer. Anal. **44**, 153–182 (2006)
20. Burger, M., Resmerita, E., He, L.: Error estimation for Bregman iterations and inverse scale space methods in image restoration. Computing **81**(2-3), 109–135 (2007)

21. Cai, J.F., Osher, S., Shen, Z.: Convergence of the linearized Bregman iteration for l1-norm minimization. Math. Comput. **78**, 2127–2136 (2009)
22. Cai, J.F., Osher, S., Shen, Z.: Linearized Bregman iterations for compressed sensing. Math. Comput. **78**, 1515–1536 (2009)
23. Cai, J.F., Osher, S., Shen, Z.: Linearized Bregman iterations for frame-based image deblurring. SIAM J. Imaging Sci. **2**, 226–252 (2009)
24. Colonius, F., Kunisch, K.: Output least squares stability in elliptic systems. Appl. Math. Optim. **19**, 33–63 (1989)
25. Combettes, P.L., Pesquet, J.C.: A proximal decomposition method for solving convex variational inverse problems. Inverse Prob. **24**, 065,014 (27 pp) (2008)
26. Daubechies, I., Defrise, M., Mol, C.D.: An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. Comm. Pure Appl. Math. **57**, 1413–1457 (2004)
27. De Cezaro, A., Haltmeier, M., Leitao, A., Scherzer, O.: On steepest-descent-Kaczmarz methods for regularizing systems of nonlinear ill-posed equations. Appl. Math. Comput. **202**, 596–607 (2008)
28. Dembo, R., Eisenstat, S., Steihaug, T.: Inexact Newton's method. SIAM J. Numer. Anal. **14**, 400–408 (1982)
29. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum Likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. B **39**, 1–38 (1977)
30. Deuflhard, P., Engl, H.W., Scherzer, O.: A convergence analysis of iterative methods for the solution of nonlinear ill-posed problems under affinely invariant conditions. Inverse Prob. **14**, 1081–1106 (1998)
31. Douglas, J., Rachford, H.H.: On the numerical solution of heat conduction problems in two and three space variables. Trans. Am. Math. Soc. **82**, 421–439 (1956)
32. Egger, H.: Fast fully iterative Newton-type methods for inverse problems. J. Inverse Ill-Posed Prob. **15**, 257–275 (2007)
33. Egger, H.: Y-Scale regularization. SIAM J. Numer. Anal. **46**, 419–436 (2008)
34. Egger, H., Neubauer, A.: Preconditioning Landweber iteration in Hilbert scales. Numer. Math. **101**, 643–662 (2005)
35. Eicke, B., Louis, A.K., Plato, R.: The instability of some gradient methods for ill-posed problems. Numer. Math. **58**, 129–134 (1990)
36. Engl, H.W., Hanke, M., Neubauer, A.: Regularization of Inverse Problems. Kluwer, Dordrecht (1996)
37. Engl, H.W., Zou, J.: A new approach to convergence rate analysis of tiknonov regularization for parameter identification in heat conduction. Inverse Prob. **16**, 1907–1923 (2000)
38. Glowinski, R., Tallec, P.L.: Augmented Lagrangian and Operator Splitting Methods in Nonlinear Mechanics. SIAM, Philadelphia (1989)
39. Green, P.J.: Bayesian reconstructions from emission tomography data using a modified EM algorithm. IEEE Trans. Med. Imaging **9**, 84–93 (1990)
40. Green, P.J.: On use of the EM algorithm for penalized likelihood estimation. J. R. Stat. Soc. Ser. B (Methodological) **52**, 443–452 (1990)
41. Haber, E.: Quasi-Newton methods for large-scale electromagnetic inverse problems. Inverse Prob. **21**, 305–323 (2005)
42. Haber, E., Ascher, U.: A multigrid method for distributed parameter estimation problems. Inverse Prob. **17**, 1847–1864 (2001)
43. Haltmeier, M., Leitao, A., Scherzer, O.: Kaczmarz methods for regularizing nonlinear ill-posed equations I: convergence analysis. Inverse Prob. Imaging **1**, 289–298 (2007)
44. Haltmeier, M., Kowar, R., Leitao, A., Scherzer, O.: Kaczmarz methods for regularizing nonlinear ill-posed equations II: applications. Inverse Prob. Imaging **1**, 507–523 (2007)
45. Haltmeier, M., Leitao, A., Resmerita, E.: On regularization methods of EM-Kaczmarz type. Inverse Prob. **25**(7), 17 (2009)
46. Hanke, M.: A regularization Levenberg-Marquardt scheme, with applications to inverse groundwater filtration problems. Inverse Prob. **13**, 79–95 (1997)

47. Hanke, M.: Regularizing properties of a truncated Newton-CG algorithm for nonlinear inverse problems. Numer. Funct. Anal. Optim. **18**, 971–993 (1997)
48. Hanke, M.: The regularizing Levenberg-Marquardt scheme is of optimal order. J. Integr. Equ. Appl. **22**, 259–283 (2010)
49. Hanke, M., Neubauer, A., Scherzer, O.: A convergence analysis of the Landweber iteration for nonlinear ill-posed problems. Numer. Math. **72**, 21–37 (1995)
50. He, L., Burger, M., Osher, S.: Iterative total variation regularization with non-quadratic fidelity. J. Math. Imaging Vis. **26**, 167–184 (2006)
51. Hein, T., Kazimierski, K.: Modified Landweber iteration in Banach spaces – convergence and convergence rates. Numer. Funct. Anal. Optim. **31**(10), 1158–1184 (2010)
52. Hochbruck, M., Hönig, M., Ostermann, A.: A convergence analysis of the exponential euler iteration for nonlinear ill-posed problems. Inverse Prob. **25**, 075,009 (18 pp) (2009)
53. Hochbruck, M., Hönig, M., Ostermann, A.: Regularization of nonlinear ill-posed problems by exponential integrators. Math. Mod. Numer. Anal. **43**, 709–720 (2009)
54. Hohage, T.: Logarithmic convergence rates of the iteratively regularized Gauß-Newton method for an inverse potential and an inverse scattering problem. Inverse Prob. **13**, 1279–1299 (1997)
55. Hohage, T.: Iterative methods in inverse obstacle scattering: regularization theory of linear and nonlinear exponentially ill-posed problems. Ph.D. thesis, University of Linz (1999)
56. Hohage, T.: Regularization of exponentially ill-posed problems. Numer. Funct. Anal. Optim. **21**, 439–464 (2000)
57. Hohage, T., Werner, F.: Iteratively regularized Newton methods with general data misfit functionals and applications to Poisson data. Numer. Math. **123**, 745–779 (2013)
58. Jin, Q.: Inexact Newton-Landweber iteration for solving nonlinear inverse problems in Banach spaces. Inverse Prob. **28**(6), 15 (2012)
59. Kaltenbacher, B.: Some Newton type methods for the regularization of nonlinear ill-posed problems. Inverse Prob. **13**, 729–753 (1997)
60. Kaltenbacher, B.: On Broyden's method for ill-posed problems. Numer. Funct. Anal. Optim. **19**, 807–833 (1998)
61. Kaltenbacher, B.: A posteriori parameter choice strategies for some Newton type methods for the regularization of nonlinear ill-posed problems. Numer. Math. **79**, 501–528 (1998)
62. Kaltenbacher, B.: A projection-regularized Newton method for nonlinear ill-posed problems and its application to parameter identification problems with finite element discretization. SIAM J. Numer. Anal. **37**, 1885–1908 (2000)
63. Kaltenbacher, B.: On the regularizing properties of a full multigrid method for ill-posed problems. Inverse Prob. **17**, 767–788 (2001)
64. Kaltenbacher, B.: Convergence rates for the iteratively regularized Landweber iteration in Banach space. In: Hömberg, D., Tröltzsch, F. (eds.) System Modeling and Optimization. 25th IFIP TC 7 Conference on System Modeling and Optimization, CSMO 2011, Berlin, Germany, September 12–16, 2011. Revised Selected Papers, pp. 38–48. Springer, Heidelberg (2013)
65. Kaltenbacher, B., Hofmann, B.: Convergence rates for the iteratively regularized Gauss-Newton method in Banach spaces. Inverse Prob. **26**, 035,007 (2010)
66. Kaltenbacher, B., Neubauer, A.: Convergence of projected iterative regularization methods for nonlinear problems with smooth solutions. Inverse Prob. **22**, 1105–1119 (2006)
67. Kaltenbacher, B., Schicho, J.: A multi-grid method with a priori and a posteriori level choice for the regularization of nonlinear ill-posed problems. Numer. Math. **93**, 77–107 (2002)
68. Kaltenbacher, B., Tomba, I.: Convergence rates for an iteratively regularized Newton-Landweber iteration in Banach space. Inverse Prob. **29**, 025010 (2013). doi:10.1088/0266-5611/29/2/025010
69. Kaltenbacher, B., Neubauer, A., Ramm, A.G.: Convergence rates of the continuous regularized Gauss-Newton method. J. Inverse Ill-Posed Prob. **10**, 261–280 (2002)
70. Kaltenbacher, B., Neubauer, A., Scherzer, O.: Iterative Regularization Methods for Nonlinear Ill-Posed Problems. Radon Series on Computational and Applied Mathematics. de Gruyter, Berlin (2008)

71. Kaltenbacher, B., Schöpfer, F., Schuster, T.: Iterative methods for nonlinear ill-posed problems in Banach spaces: convergence and applications to parameter identification problems. Inverse Prob. **25**, 065,003 (19 pp) (2009)
72. Kaltenbacher, B., Kirchner, A., Veljovic, S.: Goal oriented adaptivity in the IRGNM for parameter identification in PDEs I: reduced formulation Inverse Prob. 30, 045001 (2014)
73. Kaltenbacher, B., Kirchner, A., Vexler, B.: Goal oriented adaptivity in the IRGNM for parameter identification in PDEs II: all-at once formulations Inverse Prob. 30, 045002 (2014)
74. Kindermann, S.: Convergence analysis of minimization-based noise level-free parameter choice rules for linear ill-posed problems. ETNA **38**, 233–257 (2011)
75. Kindermann, S., Neubauer, A.: On the convergence of the quasioptimality criterion for (iterated) Tikhonov regularization. Inverse Prob. Imaging **2**, 291–299 (2008)
76. King, J.T.: Multilevel algorithms for ill-posed problems. Numer. Math. **61**, 311–334 (1992)
77. Kowar, R., Scherzer, O.: Convergence analysis of a Landweber-Kaczmarz method for solving nonlinear ill-posed problems. In: Romanov, V.G., Kabanikhin, S.I., Anikonov, Y.E., Bukhgeim, A.L. (eds.) Ill-Posed and Inverse Problems, pp. 69–90. VSP, Zeist (2002)
78. Krein, S.G., Petunin, J.I.: Scales of Banach spaces. Russ. Math. Surv. **21**, 85–160 (1966)
79. Kügler, P.: A derivative free Landweber iteration for parameter identification in certain elliptic PDEs. Inverse Prob. **19**, 1407–1426 (2003)
80. Kügler, P.: A derivative free Landweber method for parameter identification in elliptic partial differential equations with application to the manufacture of car windshields. Ph.D. thesis, Johannes Kepler University, Linz, Austria (2003)
81. Langer, S.: Preconditioned Newton methods for ill-posed problems. Ph.D. thesis, University of Göttingen (2007)
82. Langer, S., Hohage, T.: Convergence analysis of an inexact iteratively regularized Gauss-Newton method under general source conditions. J. Inverse Ill-Posed Prob. **15**, 19–35 (2007)
83. Leitao, A., Marques Alves, M.: On Landweber-Kaczmarz methods for regularizing systems of ill-posed equations in Banach spaces. Inverse Prob. **28**, 104,008 (15 pp) (2012)
84. Lions, P.L., Mercier, B.: Splitting algorithms for the sum of two nonlinear operators. SIAM J. Numer. Anal. **16**, 964–979 (1979)
85. Mülthei, H.N., Schorr, B.: On properties of the iterative maximum likelihood reconstruction method. Math. Methods Appl. Sci. **11**, 331–342 (1989)
86. Natterer, F., Wübbeling, F.: Mathematical Methods in Image Reconstruction. Society for Industrial and Applied Mathematics (SIAM), Philadelphia (2001)
87. Neubauer, A.: Tikhonov regularization of nonlinear ill-posed problems in Hilbert scales. Appl. Anal. **46**, 59–72 (1992)
88. Neubauer, A.: On Landweber iteration for nonlinear ill-posed problems in Hilbert scales. Numer. Math. **85**, 309–328 (2000)
89. Neubauer, A.: The convergence of a new heuristic parameter selection criterion for general regularization methods. Inverse Prob. **24**, 055,005 (10 pp) (2008)
90. Neubauer, A., Scherzer, O.: A convergent rate result for a steepest descent method and a minimal error method for the solution of nonlinear ill-posed problems. ZAA **14**, 369–377 (1995)
91. Osher, S., Burger, M., Goldfarb, D., Xu, J., Yin, W.: An iterative regularization method for total variation based image restoration. SIAM Multiscale Model. Simul. **4**, 460–489 (2005)
92. Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P.: Numerical Recipes: The Art of Scientific Computing, 3rd edn. Cambridge University Press, Cambridge (2007)
93. Resmerita, E., Engl, H.W., Iusem, A.N.: The expectation-maximization algorithm for ill-posed integral equations: a convergence analysis. Inverse Prob. **23**, 2575–2588 (2007)
94. Rieder, A.: On the regularization of nonlinear ill-posed problems via inexact Newton iterations. Inverse Prob. **15**, 309–327 (1999)
95. Rieder, A.: On convergence rates of inexact Newton regularizations. Numer. Math. **88**, 347–365 (2001)
96. Rieder, A.: Inexact Newton regularization using conjugate gradients as inner iteration. SIAM J. Numer. Anal. **43**, 604–622 (2005)

97. Sawatzky, A., Brune, C., Wübbeling, F., Kösters, T., Schäfers, K., Burger, M.: Accurate EM-TV algorithm in PET with low SNR. Nuclear Science Symposium Conference Record, 2008. NSS '08, pp. 5133–5137. IEEE, New York (2008)
98. Scherzer, O.: A modified Landweber iteration for solving parameter estimation problems. Appl. Math. Optim. **38**, 45–68 (1998)
99. Schöpfer, F., Louis, A.K., Schuster, T.: Nonlinear iterative methods for linear ill-posed problems in Banach spaces. Inverse Prob. **22**, 311–329 (2006)
100. Schöpfer, F., Schuster, T., Louis, A.K.: An iterative regularization method for the solution of the split feasibility problem in Banach spaces. Inverse Prob. **24**, 055,008 (20pp) (2008)
101. Schuster, T., Kaltenbacher, B., Hofmann, B., Kazimierski, K.: Regularization Methods in Banach Spaces. Radon Series on Computational and Applied Mathematics. de Gruyter, Berlin/New York (2012)
102. Stück, R., Burger, M., Hohage, T.: The iteratively regularized gauss-newton method with convex constraints and applications in 4pi-microscopy. Inverse Prob. **28**, 015,012 (2012)
103. Vardi, Y., Shepp, L.A., Kaufman, L.: A statistical model for positron emission tomography with discussion. J. Am. Stat. Assoc. **80**, 8–37 (1985)

Level Set Methods for Structural Inversion and Image Reconstruction

Oliver Dorn and Dominique Lesselier

Contents

1 Introduction 472
Level Set Methods for Inverse Problems and Image Reconstruction 472
Images and Inverse Problems 473
The Forward and the Inverse Problem 474
2 Examples and Case Studies 475
Example 1: Microwave Breast Screening 475
Example 2: History Matching in Petroleum Engineering 477
Example 3: Crack Detection 479
3 Level Set Representation of Images with Interfaces 479
The Basic Level Set Formulation for Binary Media 479
Level Set Formulations for Multivalued and Structured Media 481
Level Set Formulations for Specific Applications 485
4 Cost Functionals and Shape Evolution 489
General Considerations 489
Cost Functionals 490
Transformations and Velocity Flows 491
Eulerian Derivatives of Shape Functionals 492
The Material Derivative Method 493
Some Useful Shape Functionals 494
The Level Set Framework for Shape Evolution 495
5 Shape Evolution Driven by Geometric Constraints 497
Penalizing Total Length of Boundaries 497
Penalizing Volume or Area of Shape 499

O. Dorn (✉)
School of Mathematics, The University of Manchester, Manchester, UK

Instituto Gregorio Millán Barbany, Universidad Carlos III de Madrid, Leganés (Madrid), Spain
e-mail: oliver.dorn@manchester.ac.uk

D. Lesselier
Laboratoire des Signaux et Systemes, CNRS, Gif-sur-Yvette, France
e-mail: lesselier@lss.supelec.fr

© Springer Science+Business Media New York 2015

O. Scherzer (ed.), *Handbook of Mathematical Methods in Imaging*,
DOI 10.1007/978-1-4939-0790-8_11

6 Shape Evolution Driven by Data Misfit 499
Shape Deformation by Calculus of Variations 500
Shape Sensitivity Analysis and the Speed Method 505
Formal Shape Evolution Using the Heaviside Function 507
7 Regularization Techniques for Shape Evolution Driven by Data Misfit 512
Regularization by Smoothed Level Set Updates 512
Regularization by Explicitly Penalizing Rough Level Set Functions 514
Regularization by Smooth Velocity Fields 515
Simple Shapes and Parameterized Velocities 516
8 Miscellaneous On-Shape Evolution 516
Shape Evolution and Shape Optimization 516
Some Remarks on Numerical Shape Evolution with Level Sets 518
Speed of Convergence and Local Minima 519
Topological Derivatives 520
9 Case Studies 523
Case Study: Microwave Breast Screening 523
Case Study: History Matching in Petroleum Engineering 525
Case Study: Reconstruction of Thin Shapes (Cracks) 527
Cross-References 528
References 529

Abstract

In this chapter, an introduction is given into the use of level set techniques for inverse problems and image reconstruction. Several approaches are presented which have been developed and proposed in the literature since the publication of the original (and seminal) paper by F. Santosa in 1996 on this topic. The emphasis of this chapter, however, is not so much on providing an exhaustive overview of all ideas developed so far but on the goal of outlining the general idea of structural inversion by level sets, which means the reconstruction of complicated images with interfaces from indirectly measured data. As case studies, recent results (in 2D) from microwave breast screening, history matching in reservoir engineering, and crack detection are presented in order to demonstrate the general ideas outlined in this chapter on practically relevant and instructive examples. Various references and suggestions for further research are given as well.

1 Introduction

Level Set Methods for Inverse Problems and Image Reconstruction

The level set technique has been introduced for the solution of inverse problems in the seminal paper of Santosa [82]. Since then, it has developed significantly and appears to become now a standard technique for solving inverse problems with interfaces. However, there are still a large number of unresolved problems and open questions related to this method, which keeps fuelling active research on it worldwide. This chapter can only give a rough overview of some techniques which have been discussed so far in the literature. For more details which go beyond the material covered here, the reader is referred to the recent review articles [20, 31–

33, 92], each of them providing a slightly different view on the topic and making available a rich set of additional references which the interested reader can follow for further consultation.

Images and Inverse Problems

An *image*, as referred to in this chapter, is a (possibly vector-valued) function which assigns to each point of a given domain in 2D or in 3D one or more physical parameter values which are characteristic for that point. An image often contains *interfaces*, across which one or more of these physical parameters change value in a discontinuous manner. In many applications, these interfaces coincide with physical interfaces between different materials or regions. These interfaces divide the domain Ω in subdomains Ω_k, $k = 1, \ldots, K$ of different region-specific internal *parameter profiles*. Often, due to the different physical structures of each of these regions, quite different mathematical models might be most appropriate for describing them in the given context.

Since the image represents physical parameters, it can be tested by physical inspection. Here, the physical parameters typically appear in partial differential equations (PDEs) or in integral equations (IEs) as space-dependent coefficients, and various probing fields are created for measuring the response of the image to these inputs. Due to physical restrictions, these measurements are typically only possible at few discrete locations, often situated at the boundary of the domain Ω, but sometimes also at a small number of points inside Ω. If the underlying PDE is time dependent, then these measurements can be time-dependent functions. The corresponding measured *data* give information on the spatial distribution of the subdomains and on the corresponding internal model parameters.

Sometimes the physical interpretation of the image is that of a source distribution rather than a parameter distribution. Then, the image itself creates the probing field and needs to be determined from just one set of measured data. Also combinations are possible where some components of the (vector-valued) image describe source distributions and other components describe parameter distributions. Initial conditions or boundary conditions can also often be interpreted as images in this spirit, which need to be determined from indirect data. It is clear that this concept of an image can be generalized even further, which leads to interesting mathematical problems and concepts.

There is often a large variety of additional *prior information* available for determining the image, whose character depends on the given application. For example, it might be known or assumed that all parameter profiles inside the individual subregions of a domain Ω are constant with known or unknown region-specific values. In this particular case, only the interfaces between the different regions, and possibly the unknown parameter values, need to be reconstructed from the gathered data, which, as a mathematical problem, is much better posed [37, 71] than the task of estimating independent values at each individual pixel or voxel from the same data set without additional prior information on the image. However, in many realistic applications, the image to be found is more complicated, and even

the combined available information is not sufficient or adequate for completely and uniquely determining the underlying image. This becomes even worse due to typically noisy or incomplete data, or due to model inaccuracies. Then, it needs to be determined which information on the image is desired and which information can reasonably be expected from the data, taking into account the available additional prior information. Depending on the specific application, different viewpoints are typically taken which yield different strategies for obtaining images which agree (in an application-specific sense) with the given information. We will give some instructive examples further below.

Determining an image (or a set of possible images) from the measured data, in the above-described sense and by taking into account the available additional prior information, is called here *imaging* or *image reconstruction*. In practice, images are often represented in a computer and thereby need to be discretized somehow. The most popular discretization model uses 2D pixels or 3D voxels for representing an image, even though alternative models are possible. Often the underlying PDE also needs to be discretized on some grid, which could be done by finite differences, finite volumes, finite elements, and other techniques. The discretization for the image does not necessarily need to be identical to the discretization used for solving the PDE, and sometimes different models are used for discretizing the image and the PDE. However, in these cases, some method needs to be provided to map from one representation to the other. In a level set representation of an image, also the level set functions need to be discretized for being represented in a computer. The above said then holds true also for the discretizations of the level set functions, which could either follow the same model as the PDE and/or a pixel model for the image or follow a different pattern.

The Forward and the Inverse Problem

In this chapter, it is supposed that data $\tilde{g}$ are given in the form

$$\tilde{g} = \mathscr{M}\tilde{\mathbf{u}}, \tag{1}$$

where $\mathscr{M}$ denotes a linear measurement operator, and $\tilde{\mathbf{u}}$ are the physical states created by the sources $\mathbf{q}$ for probing the image. It is assumed that a physical model $\Lambda(b)$ is given, which incorporates the (possibly vector-valued) model parameter b and which is able to (roughly) predict the probing physical states when being plugged into an appropriate numerical simulator, provided the correct sources and physical parameters during the measurement process were known. The forward operator $\mathscr{A}$ is defined as

$$\mathscr{A}(b, \mathbf{q}) = \mathscr{M}\Lambda(b)^{-1}\mathbf{q}. \tag{2}$$

As mentioned, $\Lambda(b)$ is often described in a form of some partial differential equation (PDE) or, alternatively, an integral equation (IE), and the individual coefficients of the model parameter b appear at one or several places in this model

as space-dependent coefficients. In most applications, measurements are taken only at few locations of the domain, for example, at the boundary of the area of interest, from which the physical parameters b or the source $\mathbf{q}$ (or both) need to be inferred in the whole domain. It is said that with respect to these unknowns, the measurements are indirect: They are taken not at the locations where the unknowns need to be determined but indirectly by their overall impact on the states (modeled by the underlying PDE or IE) probing the image, which are measured only at few locations. The behavior of the states is modeled by the operator $\mathscr{A}$ in (2). If in $\mathscr{A}$ only b (but not $\mathbf{q}$) is unknown, then the problem is an inverse parameter or inverse scattering problem. If in $\mathscr{A}$ only $\mathbf{q}$ (but not b) is unknown, then the problem is an inverse source problem. Given measured data $\tilde{g}$, the "residual operators" $\mathscr{R}$ are correspondingly given by

$$\mathscr{R}(b, \mathbf{q}) = \mathscr{A}(b, \mathbf{q}) - \tilde{g}. \tag{3}$$

Given the above definitions, an *image* is defined here as a mapping

$$a : \Omega \to \mathbb{R}^n,$$

where Ω is a bounded or unbounded region in $\mathbb{R}^2$ or in $\mathbb{R}^3$ and n is the number of components of the (vector-valued) image. Each component function a_k, $k = 1, \ldots, n$, represents a space-dependent physical characteristic of the domain Ω which can be probed by physical inspection. If it appears as a coefficient of a PDE (or IE), it is denoted $a_k = b_k$, and if it appears as a source, it is denoted $a_k = q_k$. The exposition given in this chapter mainly focuses on the recovery of parameter distributions $a_k = b_k$ and addresses several peculiarities related to those cases. However, the main concepts carry over without major changes to inverse source problems and also to some related formulations as, for example, the reconstruction of boundary or initial conditions of PDEs.

2 Examples and Case Studies

Some illustrative examples and case studies are presented in the following, which will be used further on in this chapter for demonstrating basic ideas and concepts on realistic and practical situations.

Example 1: Microwave Breast Screening

Figure 1 shows two-dimensional images from the application of microwave breast screening. The images of size 160×160 pixels have been constructed synthetically based on MRI images of the female breast. Three representative breast structures are displayed in the three images of the left column, where the value at each pixel of the images represents the physical parameter "static relative permittivity."

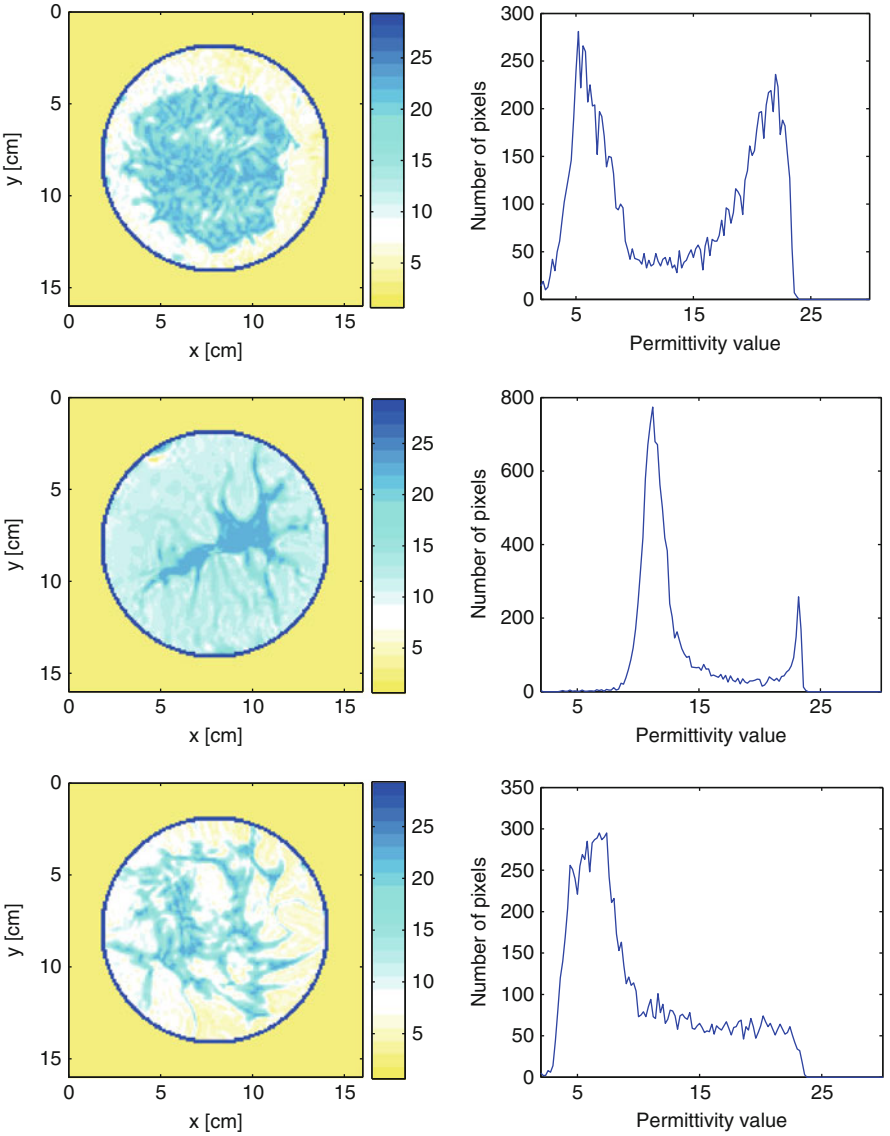


Fig. 1 Three images from microwave breast screening. The three images are synthetically generated from MRI breast models. *Left column*: two-dimensional maps of the distribution of the static permittivity ε_{st} inside the three breast models. *Right column*: the corresponding histograms of values of ε_{st} in each map

A commonly accepted model for breast tissue is to roughly distinguish between skin, fatty tissue and fibroglandular tissue. In the images also a matching liquid is shown in which the breast is immersed. Inside the breast, regions can be identified easily which correspond to fibroglandular tissue (high static relative permittivity

values) and fatty tissue (low static relative permittivity values), separated by a more or less complicated interface. On the right column, histograms are shown for the distributions of static relative permittivity values inside the breast. In these histograms, it becomes apparent that values for fatty and fibroglandular tissue are clustered around two typical values, but with a broader range of distribution. However, a clear identification of fatty and fibroglandular tissue cannot be made easily for each pixel of the image based on just these values.

Nevertheless, during a reconstruction, and from anatomical reasoning, it does make sense to assume a model where fatty and fibroglandular tissue occupy some subregions of the breast where a sharp interface exists between these subregions. Finding these subregions provides valuable information for the physician. Furthermore, it might be sufficient for an overall evaluation of the situation to have a smoothly varying profile of tissue parameters reconstructed inside each of these subregions, allowing for the choice of a smoothly varying profile of static relative permittivity values inside each region. In the same spirit, from anatomical reasoning, it makes sense to assume a sharp interface (now of less complicated behavior) separating the skin region from the fatty/fibroglandular tissue on the one side and from the matching liquid on the other side. It might also be reasonable to assume that the skin and the matching liquid have constant static permittivity values, which might be known or not. If a tumor in its early stage of development is sought in this breast model, it will occupy an additional region of small size (and either simple or complicated shape and topology) and might have constant but unknown static relative permittivity value inside this region.

During a reconstruction for breast screening, this set of plausible assumptions provides us with a complex mathematical breast model which incorporates this prior information and might yield an improved and more realistic image for the reconstructed breast (including a better estimate of the tumor characteristics) than a regular pixel-based inversion would be able to provide. This is so because it is assumed that the real breast follows roughly the complicated model constructed above and that this additional information is taken into account in the inversion.

In this application, the underlying PDE is the system of time-harmonic Maxwell's equations, or its 2D representative (describing so-called TM-waves), a Helmholtz equation. The "static relative permittivity," as mapped in Fig. 1, represents one parameter entering in the wavenumber of the Debye dispersion model. The electromagnetic fields are created by specifically developed microwave antennas surrounding the breast, and the data are gathered at different microwave antennas also located around the breast. For more details, see [52].

Example 2: History Matching in Petroleum Engineering

Figure 2 shows a synthetically created 2D image of a hydrocarbon reservoir during the production process. Circles indicate injection wells, and crosses indicate production wells. The physical parameter displayed in the image is the permeability, which affects fluid flow in the reservoir. Physically, two lithofacies can be distinguished in this image, namely, sandstone and shaly sandstone (further on simply

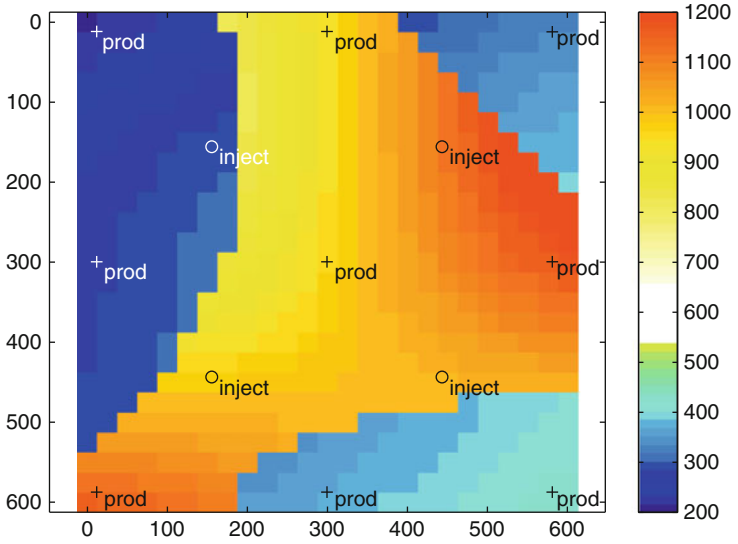


Fig. 2 An image from reservoir engineering. Shown is the permeability distribution of a fluid flow model in a reservoir which consists of a sandstone lithofacie (values in the range of 150–500 mDarcy) and a shaly sandstone lithofacie (values in the range of 900–1,300 mDarcy), separated by a sharp interface. The sandstone region shows an overall linear trend in the permeability distribution, whereas the shaly sandstone region does not show any clear trend

called "shale"). The sandstone region has permeability values roughly in the range 150–500 mDarcy, whereas shale has permeability values more in the range 900–1,300 mDarcy. In petroleum engineering applications, the parameters inside a given lithofacie sometimes follow an overall linear trend, which is the case here inside the sandstone region. This information is often available from geological evaluation of the terrain. As a rough approximation, inside this region, the permeability distribution can be modeled mathematically as a smooth perturbation of a bilinear model. Inside the shale region, no trend is observed or expected, and therefore the permeability distribution is described as a smooth perturbation of a constant distribution (i.e., an overall smoothly varying profile).

During a reconstruction, a possible model would be to reconstruct a reservoir image from production data which consists of three different quantities: (1) the interface between the sandstone and shale lithofacies, (2) the smooth perturbation of the constant profile inside the shale region, and (3) the overall trend (i.e., the bilinear profile) inside the sandstone region, assuming that inside this sandstone region the smooth perturbation is small compared to this dominant overall trend. In this application, the PDE is a system of equations modeling two-phase or three-phase fluid flow in a porous medium, of which the relative permeability is one model parameter.

The "fields" (in a slightly generalized sense) are represented in this application by pressure values and water/oil saturation values at each point inside the reservoir

during production and are generated by injecting (under high pressure) water in the injection wells and extracting (imposing lower pressure) water and oil from the production wells. The data are the injection and production rates of water and oil, respectively, and sometimes pressure values measured at injection and production wells over production time. For more details, see [35].

Example 3: Crack Detection

Figure 3 shows an image of a disconnected crack embedded in a homogeneous material. The cracks are represented in this simplified model as very thin regions of fixed thickness. The physical parameter represented by the image is the conductivity distribution in the domain. Only two values can be assumed by this conductivity, one inside the thin region (crack) and another one in the background. The background value is typically known, and the value inside the crack might either be approximately known or it might be an unknown of the inverse problem. The same holds true for the thickness of the crack, which is assumed constant along the cracks, even though the correct thickness (the constant) might become an unknown of the inverse problem as well. Here insulating cracks are considered, where the conductivity is significantly lower than in the background. The probing fields inside the domain are the electrostatic potentials which are produced by applying voltages at various locations along the boundary of the domain, and the data are the corresponding currents across the boundary at discrete positions.

This model can be considered as a special case of a binary medium where volumetric inclusions are embedded in a homogeneous background. However, the fact that these structures are very thin with fixed thickness requires some special treatment during the shape evolution, which will be commented on further below. In this application, the underlying PDE is a second-order elliptic equation modeling the distribution of electric potentials in the domain for a set of given applied voltage patterns. For more details, see [4].

3 Level Set Representation of Images with Interfaces

A complex image in the above sense needs a convenient mathematical representation in order to be dealt with in a computational and mathematical framework. In this section, several different approaches are listed which have been proposed in the literature for describing images with interfaces by a level set technique. First, the most basic representation is given, which only considers binary media. Afterwards, various representations are described which represent more complicated situations.

The Basic Level Set Formulation for Binary Media

In the shape inverse problem in its simplest form, the parameter distribution is described by

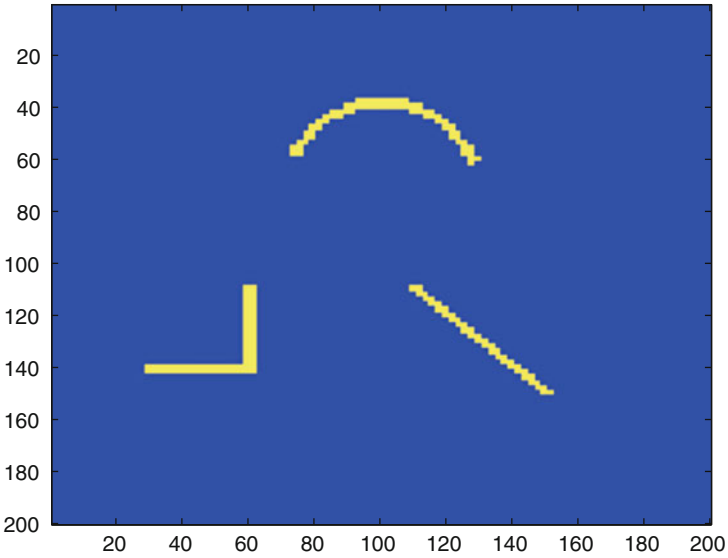


Fig. 3 An image from the application of crack detection. Three disconnected crack components are embedded in a homogeneous background medium and need to be reconstructed from electrostatic measurements at the region boundary. In the considered case of insulating cracks, these components are modeled as thin shapes of fixed thickness with a conductivity value much lower than the background conductivity

$$b(\mathbf{x}) = \begin{cases} b^{(i)}(\mathbf{x}) & \text{in} \quad D \\ b^{(e)}(\mathbf{x}) & \text{in} \quad \Omega \backslash D \end{cases}, \tag{4}$$

where $D \subset \Omega$ is a subregion of Ω and where usually discontinuities in the parameters b occur at the interface ∂D. In the *basic level set representation for the shape* D, a (sufficiently smooth, i.e., Lipschitz continuous) level set function $\phi : \Omega \to \mathbb{R}$ is introduced and the shape D is described by

$$\begin{cases} \phi(\mathbf{x}) \leq 0 & \text{for all} \quad \mathbf{x} \in D, \\ \phi(\mathbf{x}) > 0 & \text{for all} \quad \mathbf{x} \in \Omega \backslash D. \end{cases} \tag{5}$$

In other words, the parameter function b has the form

$$b(\mathbf{x}) = \begin{cases} b^{(i)}(\mathbf{x}) & \text{where} \quad \phi(\mathbf{x}) \leq 0 \\ b^{(e)}(\mathbf{x}) & \text{where} \quad \phi(\mathbf{x}) > 0. \end{cases} \tag{6}$$

Certainly, a unique representation of the image is possible by just knowing those points where $\phi(\mathbf{x})$ has a change of sign (the so-called zero level set) and additionally knowing the two interior profiles $b^{(i)}(\mathbf{x})$ and $b^{(e)}(\mathbf{x})$ inside those areas of Ω where they are active (which are D and $\Omega \backslash D$, respectively). Often, however, it is more

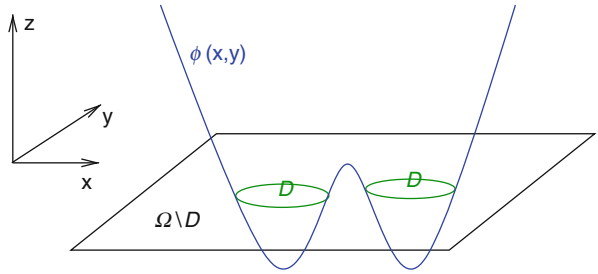


Fig. 4 The basic level set representation of a shape D. Those points of the domain where the describing level set function assumes negative values are "inside" the shape D described by the level set function ϕ, while those with positive values are "outside" it. The zero level set where $\phi = 0$ represents the shape boundary

convenient to assume that these functions are defined on larger sets which include the minimal sets mentioned above. In this chapter, it is assumed that all functions are defined on the entire domain Ω, by employing any convenient extensions from the abovementioned sets to the rest of Ω. Again, it is clear that the above extensions are not unique and that many possible representations can then be found for a given image. Which one to choose depends on details of the algorithm for constructing the image, on the available prior information, and possibly on other criteria (Fig. 4).

For a sufficiently smooth level set function, the boundary of the shape D permits the characterization

$$\partial D = \{\mathbf{x} \in \Omega, \quad \phi(\mathbf{x}) = 0\}. \tag{7}$$

This representation motivates the name *zero level set* for the boundary of the shape. In some representations listed further below, however, level set functions are preferred which are discontinuous across those sets where they change sign. Then, the boundary of the different regions can be defined alternatively as

$$\begin{aligned} \partial D = \{\mathbf{x} \in \Omega \ : \ &\text{for all } \rho > 0 \text{ we can find } \mathbf{x}_1, \mathbf{x}_2 \in B_\rho(\mathbf{x}) \\ &\text{with } \phi(\mathbf{x}_1) > 0 \text{ and } \phi(\mathbf{x}_2) \leq 0\} \end{aligned} \tag{8}$$

where $B_\rho(\mathbf{x}_0) = \{\mathbf{x} \in \Omega : |\mathbf{x} - \mathbf{x}_0| < \rho\}$.

Level Set Formulations for Multivalued and Structured Media

As mentioned already above, in many applications the binary model described in section "The Basic Level Set Formulation for Binary Media" is not sufficient and more complex image models need to be employed. Several means have been discussed in the literature for generalizing the basic model to more complex situations, some of them being listed in the following.

Different Levels of a Single Smooth Level Set Function

A straightforward generalization of the technique described in section "The Basic Level Set Formulation for Binary Media" consists in using, in addition to the level set zero, additional level sets of a given smooth (e.g., Lipschitz continuous) level set function in order to describe different regions of a given domain [99]. For example, define

$$\Gamma_i = \{\mathbf{x} \in \Omega, \quad \phi(\mathbf{x}) = c_i\} \tag{9}$$

$$D_i = \{\mathbf{x} \in \Omega, \quad c_{i+1} < \phi(\mathbf{x}) < c_i\}, \tag{10}$$

where c_i are prespecified values with $c_{i+1} > c_i$ for $i = 0, \ldots, \underline{i} - 1$ and with $c_0 = +\infty$, $c_{\underline{i}} = -\infty$. Then,

$$\Omega = \bigcup_{i=0}^{\underline{i}} D_i, \quad \text{with} \quad D_i \cap D_{i'} = \emptyset \quad \text{for } i \neq i'. \tag{11}$$

A *level set representation for the image* b is then given as a tupel $(b_0, \ldots, b_{\underline{i}}, \phi)$ which satisfies

$$b(\mathbf{x}) = b_i(\mathbf{x}) \text{ for } c_{i+1} < \phi(\mathbf{x}) < c_i. \tag{12}$$

It is clear that certain topological restrictions are imposed on the distribution of the regions D_i by this formulation. In particular, it favors certain nested structures. For more details, see [63].

Piecewise Constant Level Set Function

This model describes piecewise constant multiple phases of a domain by only one level set function and has its origins in the application of image segmentation. A single level set function is used which is only allowed to take a small number of different values, e.g.,

$$\phi(\mathbf{x}) = i \quad \text{in } D_i, \quad \text{for } i = 0, \ldots, \underline{i}, \tag{13}$$

$$\Omega = \bigcup_{i=0}^{\underline{i}} D_i, \quad \text{with} \quad D_i \cap D_{i'} = \emptyset \quad \text{for} \quad i \neq i'.$$

Introducing the set of basis functions γ_i

$$\gamma_i = \frac{1}{\alpha_i} \prod_{\substack{j=1 \\ j \neq i}}^{\underline{i}} (\phi - j) \quad \text{with} \quad \alpha_i = \prod_{\substack{j=1 \\ j \neq i}}^{\underline{i}} (i - j), \tag{14}$$

the parameter distribution $b(\mathbf{x})$ is defined as

$$b = \sum_{i=1}^{\underline{i}} b_i \gamma_i. \tag{15}$$

A *level set representation for the image* b is then given as a tupel $(b_1, \ldots, b_{\underline{i}}, \phi)$ with

$$b(\mathbf{x}) = b_i \quad \text{where } \phi(\mathbf{x}) = i. \tag{16}$$

Numerical results using this model can be found, among others, in [61, 65, 67, 97].

Vector Level Set

In [99] multiple phases are described by using one individual level set function for each of these phases, i.e.,

$$\Gamma_i = \{\mathbf{x} \in \Omega, \quad \phi_i(\mathbf{x}) = 0\} \tag{17}$$

$$D_i = \{\mathbf{x} \in \Omega, \quad \phi_i(\mathbf{x}) \leq 0\}, \tag{18}$$

for sufficiently smooth level set functions ϕ_i, $i = 0, \ldots, \underline{i}$. In this model, the *level set representation for the image* b is given by a tupel $(b_1, \ldots, b_{\underline{i}}, \phi_1, \ldots, \phi_{\underline{i}})$ which satisfies

$$b(x) = b_k(x) \text{ where } \phi_k(\mathbf{x}) \leq 0. \tag{19}$$

Care needs to be taken here that different phases do not overlap, which is not automatically incorporated in the model. For more details on how to address this and other related issues, see [99].

Color Level Set

An alternative way of describing different phases by more than one level set functions has been introduced in [95] in the framework of image segmentation and further investigated by [14, 25, 38, 63, 92] in the framework of inverse problems. In this model (which also is known as the Chan–Vese model), up to 2^n different phases can be represented by n different level set functions by distinguishing all possible sign combinations for these functions. For example, a *level set representation for an image* b *containing up to four different phases* is given by the tupel $(b_1, b_2, b_3, b_4, \phi_1, \phi_2)$ which satisfies

$$\begin{aligned} b(\mathbf{x}) = {} & b_1(1 - H(\phi_1))(1 - H(\phi_2)) + b_2(1 - H(\phi_1))H(\phi_2) \\ & + b_3 H(\phi_1)(1 - H(\phi_2)) + b_4 H(\phi_1)H(\phi_2). \end{aligned} \tag{20}$$

Also here, the contrast values b_ν, $\nu = 1, \ldots, 4$ are allowed to be smoothly varying functions inside each region. The four different regions are then given by

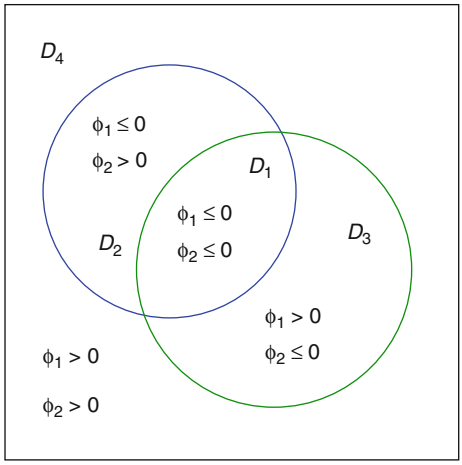


Fig. 5 Color level set representation of multiple shapes. Each region is characterized by a different sign combination of the two describing level set functions

$$
\begin{aligned}
D_1 &= \{\mathbf{x}, \quad \phi_1 \leq 0 \quad \text{and} \quad \phi_2 \leq 0\} \\
D_2 &= \{\mathbf{x}, \quad \phi_1 \leq 0 \quad \text{and} \quad \phi_2 > 0\} \\
D_3 &= \{\mathbf{x}, \quad \phi_1 > 0 \quad \text{and} \quad \phi_2 \leq 0\} \\
D_4 &= \{\mathbf{x}, \quad \phi_1 > 0 \quad \text{and} \quad \phi_2 > 0\}.
\end{aligned} \tag{21}
$$

This yields a complete covering of the domain Ω by the four regions, each point $\mathbf{x} \in \Omega$ being part of exactly one of the four shapes D_ν; see Fig. 5.

Binary Color Level Set

An alternative technique for using more than one level set function for describing multiple phases, which is, in a certain sense, a combination of the piecewise constant level set model described in section "Piecewise Constant Level Set Function" and the color level set technique described in section "Color Level Set," has been proposed in [62] for the application of Mumford–Shah image segmentation. For the description of up to four phases by two (now piecewise constant) level set functions ϕ_1 and ϕ_2, in this binary level set model, the two level set functions are required to satisfy

$$
\phi_i \in \{-1, 1\}, \quad \text{or} \quad \phi_i^2 = 1, \quad i \in \{1, 2\}. \tag{22}
$$

The parameter function $b(\mathbf{x})$ is given by

$$
\begin{aligned}
b(\mathbf{x}) = \frac{1}{4}\Big(& b_1(\phi_1 - 1)(\phi_2 - 1) - b_2(\phi_1 - 1)(\phi_2 + 1) \\
& - b_3(\phi_1 + 1))\,(\phi_2 - 1) + b_4(\phi_1 + 1)(\phi_2 + 1)\Big),
\end{aligned} \tag{23}
$$

and the four different regions are encoded as

$$
\begin{aligned}
D_1 &= \{\mathbf{x}, \quad \phi_1 = -1 \quad \text{and} \quad \phi_2 = -1\} \\
D_2 &= \{\mathbf{x}, \quad \phi_1 = -1 \quad \text{and} \quad \phi_2 = +1\} \\
D_3 &= \{\mathbf{x}, \quad \phi_1 = +1 \quad \text{and} \quad \phi_2 = -1\} \\
D_4 &= \{\mathbf{x}, \quad \phi_1 = +1 \quad \text{and} \quad \phi_2 = +1\}.
\end{aligned} \tag{24}
$$

A *level set representation for an image b containing up to four different phases* is given by the tupel $(b_1, b_2, b_3, b_4, \phi_1, \phi_2)$ which satisfies (23). For more details, we refer to [62].

Level Set Formulations for Specific Applications

Often, for specific applications, it is convenient to develop particular modifications or generalizations of the above-described general approaches for describing multiple regions by taking into account assumptions and prior information which are very specific to the particular application. A few examples are given below.

A Modification of Color Level Set for Tumor Detection

In the application of tumor detection from microwave data for breast screening (see section "Example 1: Microwave Breast Screening"), the following situation needs to be modeled. The breast consists of four possible tissue types, namely, the skin, fibroglandular tissue, fatty tissue, and a possible tumor. Each of these tissue types might have an internal structure, which is (together with the mutual interfaces) one unknown of the inverse problem. In principle, the color level set description using two level set functions for describing four different phases would be sufficient for modeling this situation. However, the reconstruction algorithm as presented in [52] requires some flexibility with handling these four regions separately, which is difficult in this minimal representation of four regions. Therefore, in [52], the following modified version of the general representation of color level sets is proposed for modeling this situation. In this modified version, m different phases (here $m = 4$) are described by $n = m - 1$ level set functions in the following form

$$
\begin{aligned}
b(\mathbf{x}) = b_1(1 - H(\phi_1)) + H(\phi_1)\Big[& b_2(1 - H(\phi_2)) \\
& + H(\phi_2)\left\{b_3(1 - H(\phi_3)) + b_4 H(\phi_3)\right\}\Big]
\end{aligned} \tag{25}
$$

or

$$
\begin{aligned}
D_1 &= \{\mathbf{x}, \quad \phi_1 \le 0\} \\
D_2 &= \{\mathbf{x}, \quad \phi_1 > 0 \quad \text{and} \quad \phi_2 \le 0\} \\
D_3 &= \{\mathbf{x}, \quad \phi_1 > 0 \quad \text{and} \quad \phi_2 > 0 \quad \text{and} \quad \phi_3 \le 0\} \\
D_4 &= \{\mathbf{x}, \quad \phi_1 > 0 \quad \text{and} \quad \phi_2 > 0 \quad \text{and} \quad \phi_3 > 0\},
\end{aligned} \tag{26}
$$

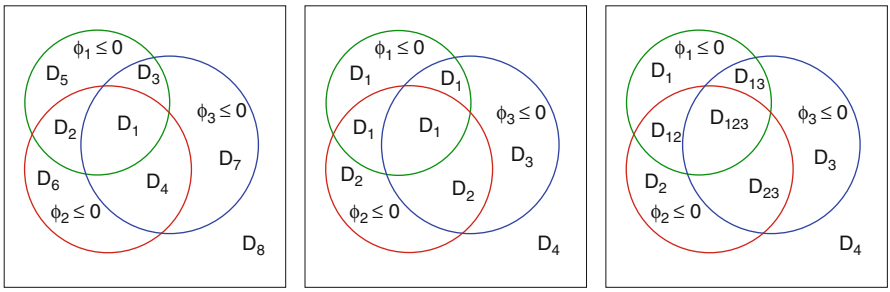


Fig. 6 Multiple level set representation for modeling multiphase inverse problems. *Left*: original color level set technique for describing eight different phases by the different sign combinations of three level set functions. *Center*: modified color level set technique used in the model for early detection of breast cancer from microwave data. The possible eight regions of the color level set presentation are filled with four different materials in a tailor-made fashion for this application. *Right*: modified color level set technique for modeling the history matching problem of a water-flooding process in a petroleum reservoir. Also here the eight different regions are filled by a specific combination of materials characteristic for the reconstruction scheme used in this application. Regions with more than one subindex correspond to "characteristic regions" with averaged parameter values

where b_1, b_2, b_3, and b_4 denote the dielectric parameters of the skin, tumorous, fibroglandular, and fatty tissue, respectively. In (25), ϕ_1, ϕ_2, and ϕ_3 are the three different level set functions indicating the regions filled with the skin, tumorous, and fibroglandular tissue, respectively, and the contrast values b_ν, $\nu = 1, \ldots, 4$ are generally allowed to be smoothly varying functions inside each region. This combination of $m - 1$ level set functions for describing m different phases has certain advantages with respect to the standard color level set formulation during the reconstruction process, as it is pointed out in [52]. On the other hand, it is obvious that (26) can be considered a special case of the color level set technique (section "Color Level Set") where the theoretically possible $2^3 = 8$ different values of the color level set description are enforced to fall into $m = 4$ different groups of characteristic values; see the central image of Fig. 6.

A Modification of Color Level Set for Reservoir Characterization

Another modification of the color level set technique has been used in [35] for the application of history matching in reservoir engineering; see section "Example 2: History Matching in Petroleum Engineering." Given, as an example, $n = 4$ level set functions $\phi_1, \ldots, \phi_4$, we define the parameter (permeability) distribution inside the reservoir by

$$\begin{aligned} b = {} & b_1(1 - H(\phi_1))H(\phi_2)H(\phi_3) + b_2 H(\phi_1)(1 - H(\phi_2))H(\phi_3) \\ & + b_3 H(\phi_1)H(\phi_2)(1 - H(\phi_3)) + b_4 H(\phi_1)H(\phi_2)H(\phi_3) \\ & + \frac{b_2 + b_3}{2} H(\phi_1)(1 - H(\phi_2))(1 - H(\phi_3)) \end{aligned}$$

$$+\frac{b_1+b_3}{2}(1-H(\phi_1))H(\phi_2)(1-H(\phi_3))$$
$$+\frac{b_1+b_2}{2}(1-H(\phi_1))(1-H(\phi_2))H(\phi_3)$$
$$+\frac{b_1+b_2+b_3}{3}(1-H(\phi_1))(1-H(\phi_2))(1-H(\phi_3)), \tag{27}$$

where the permeability values b_ν, $\nu = 1,\ldots,4$ are assumed constant inside each region. The four lithofacies are represented as

$$\begin{aligned} D_1 &= \{\mathbf{x}, \quad \phi_1 \le 0 \quad \text{and} \quad \phi_2 > 0 \quad \text{and} \quad \phi_3 > 0\} \\ D_2 &= \{\mathbf{x}, \quad \phi_2 \le 0 \quad \text{and} \quad \phi_3 > 0 \quad \text{and} \quad \phi_1 > 0\} \\ D_3 &= \{\mathbf{x}, \quad \phi_3 \le 0 \quad \text{and} \quad \phi_1 > 0 \quad \text{and} \quad \phi_2 > 0\} \\ D_4 &= \{\mathbf{x}, \quad \phi_1 > 0 \quad \text{and} \quad \phi_2 > 0 \quad \text{and} \quad \phi_3 > 0\}. \end{aligned} \tag{28}$$

Let in the following $n = 4$ be the number of lithofacies. In this model, a point in the reservoir corresponds to the lithofacie D_l, ($l = 1,\ldots,n-1$) if ϕ_l has negative sign and all the other level set functions have positive sign. In addition, one lithofacie (which here is referred to as the "background" lithofacie with index $l = n$) corresponds to those points where none of the level set functions has a negative sign. Notice that typically this definition does not yield a complete covering of the whole domain Ω by the four (n) lithofacies; see the right image of Fig. 6. Those regions inside the domain where more than one level set function are negative are recognized as so-called critical regions and are introduced for providing a smooth evolution from the initial guess to the final reconstruction. Inside these critical regions, the permeability assumes values which are calculated as certain averages over values of the neighboring noncritical regions. They are indicated in the right image of Fig. 6 by using multiple subindices indicating which noncritical regions contribute to this averaging procedure. For more details regarding this model, and numerical experiments for the application of reservoir characterization, see [35].

A Modification of the Classical Level Set Technique for Describing Cracks or Thin Shapes

Cracks of finite thickness can be modeled by using two level set functions in a setup which amounts to a modification of the classical level set technique for binary media. For simplicity, assume that a crack or thin region of finite thickness is embedded in a homogeneous background. The classical level set technique described in section "The Basic Level Set Formulation for Binary Media" captures this situation in principle, since the crack can be interpreted as a simple shape (with possibly complicated topology) embedded in a homogeneous background. However, when it comes to shape evolution for such a crack-like structure, it is difficult to maintain a fixed thickness of the thin shape following the classical

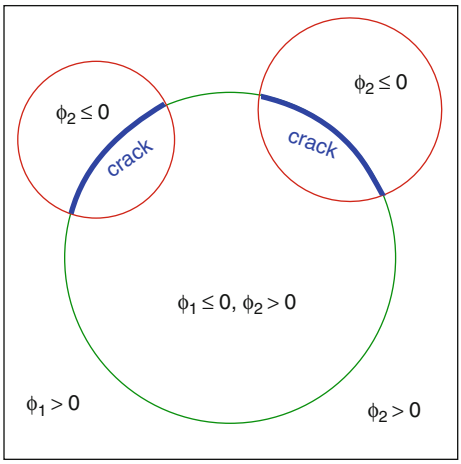


Fig. 7 Multiple level set representation for modeling a disconnected crack. The zero level set of the first level set function defines the potential outline of the crack, of which the second level set function selects those parts that are actually occupied by the crack. The "ideal" crack has vanishing thickness, whereas the "real" crack modeled by the level set technique has a small finite thickness and is practically obtained by a narrow band technique

shape evolution scheme. This is so since the classical shape evolution applies an individually calculated velocity field value in the normal direction at each point of the entire shape boundary, such that the thickness of the thin region will not be maintained. For crack evolution, the deformations of adjacent boundary points need to be coordinated in order to maintain the thickness of the crack during the entire shape evolution; see Fig. 9.

A modified version of the classical level set technique has been proposed in [4, 77] which uses two level set functions for modeling crack propagation and crack reconstruction in this sense. Here, a small neighborhood (narrowband) of the zero level set of the first level set function defines the general outline of the crack, whereas the second level set function selects those parts of this band structure which in fact contribute to the possibly disconnected crack topology.

In more details, given a continuously differentiable level set function ϕ_1 and its zero level set

$$\Gamma_{\phi_1} = \{\mathbf{x} \in \Omega \ : \ \phi_1(\mathbf{x}) = 0\}. \tag{29}$$

The normal $\mathbf{n}$ to Γ_{ϕ_1} is given by (61) and is pointing into the direction where $\phi(\mathbf{x}) \geq 0$. An (connected or disconnected) "ideal" (i.e., of thickness zero) crack with finite length completely contained inside Ω is constructed by introducing a second level set function ϕ_2 which selects one or more parts from the line Γ_{ϕ_1}; see Fig. 7. This second level set function defines the region

$$B = \{\mathbf{x} \in \Omega \ : \ \phi_2(\mathbf{x}) \leq 0\}. \tag{30}$$

The "ideal" crack is then defined as a collection of finite subintervals of Γ_{ϕ_1}

$$S[\phi_1, \phi_2] = \Gamma_{\phi_1} \cap B. \tag{31}$$

An ideal "insulating" crack S (of thickness zero) is then supplemented with a vanishing electrical current condition across this set $S[\phi_1, \phi_2]$. However, in the simplified level set model, not the ideal crack is considered, but cracks of finite thickness $2\delta > 0$ with known conductivity b_i inside the crack and b_e outside it. Moreover, in the insulating case, it is assumed that $b_i \ll b_e$. In this model, a small neighborhood of Γ_{ϕ_1} is introduced as

$$\Gamma_{\phi_1}^{\delta} = \{\mathbf{y} \in \Omega : \mathbf{y} = \mathbf{x} - \tau \mathbf{n}(\mathbf{x}), |\tau| < \delta, \mathbf{x} \in \Gamma_{\phi_1}\}, \tag{32}$$

and the above-defined "ideal crack" S is associated now with a "real crack" counterpart

$$S_{\delta} = \Gamma_{\phi_1}^{\delta} \cap B. \tag{33}$$

The conductivity distribution is

$$b(\mathbf{x}) = \begin{cases} b_i & \text{for} \quad \mathbf{x} \in S_{\delta} \\ b_e & \text{otherwise} \end{cases} \tag{34}$$

in the domain Ω. Certainly, the real crack can also alternatively be defined by

$$\tilde{S}_{\delta} = \{\mathbf{y} \in \Omega : \mathbf{y} = \mathbf{x} - \tau \mathbf{n}(\mathbf{x}),\ |\tau| < \delta,\ \mathbf{x} \in S\}, \tag{35}$$

which would slightly change the shape of the crack at the crack tips. Here, the form (33) is preferred. For the numerical treatment, see [4, 77].

4 Cost Functionals and Shape Evolution

One important technique for creating images with interfaces satisfying certain criteria is *shape evolution*, more specifically, *interface and profile evolution*. The general goal is to start with a set of shapes and profiles as initial guess, and then let both, shapes and profiles, evolve due to some appropriate evolution laws in order to improve the initial guess with increasing artificial evolution time. The focus in the following will be on shape evolution, since evolution laws for interior profiles fairly much follow classical and well-known concepts. Evolution of a shape or an interface can be achieved either by defining a velocity field on the domain Ω which deforms the boundaries of this shape or by defining evolution laws directly for the level set functions representing the shape. Some of these techniques will be presented next.

General Considerations

In many applications, images need to be evaluated for verifying their usefulness or *merit* for a given situation. This evaluation is usually based on a number of

criteria, among them being the ability of the image (in its correct interpretation) to reproduce the physically measured data (its *data fitness*). Other criteria include the consistence with any additionally available prior knowledge on the given situation or the closeness of the image to a set of reference images. In many cases, some form of *merit function* (often in terms of a properly defined *cost functional*) is defined whose value is intended to indicate the usefulness of the image in a given application. However, sometimes this decision is done based on visual inspection only.

In general, during this evaluation process, a family of images is created and the merit of each of these images is assessed. Then, one or more of these images are selected. Let $\left(b^{(1)}, \ldots, b^{(\underline{i})}, \phi^{(1)}, \ldots, \phi^{(\underline{j})}\right)$ be a level set representation for the class of images to be considered. Then, creating this family of images can be described either in a continuous way by an artificial time evolution

$$\left(b^{(1)}(t), \ldots, b^{(\underline{i})}(t), \phi^{(1)}(t), \ldots, \phi^{(\underline{j})}(t)\right), \quad t \in [0, t_{\max}],$$

with an artificial evolution time t or in a discrete way

$$\left(b_k^{(1)}, \ldots, b_k^{(\underline{i})}, \phi_k^{(1)}, \ldots, \phi_k^{(\underline{j})}\right), \quad k = 1, \ldots, \underline{k},$$

with a counting index k. Usually these images are created in a sequential manner, using evolution laws

$$\frac{d}{dt}\left(b^{(1)}(t), \ldots, b^{(\underline{i})}(t), \phi^{(1)}(t), \ldots, \phi^{(\underline{j})}(t)\right) = f(t),$$

with a multicomponent forcing term $f(t)$ or update formulas

$$\left(b_{k+1}^{(1)}, \ldots, b_{k+1}^{(\underline{i})}, \phi_{k+1}^{(1)}, \ldots, \phi_{k+1}^{(\underline{j})}\right) = F_k\left(b_k^{(1)}, \ldots, b_k^{(\underline{i})}, \phi_k^{(1)}, \ldots, \phi_k^{(\underline{j})}\right)$$

with update operators F_k. These evolution laws and update operators can also be defined on ensembles of images, which allows for statistical evaluation of each ensemble during the evaluation process. Any arbitrarily defined evolution law and set of update operators yield a family of images which can be evaluated, but typically those are preferred which point into a descent direction of some predefined *cost functional*. Some choices of such cost functionals will be discussed in the following.

Cost Functionals

In general, a cost functional can consist of various components, typically combined in an additive or multiplicative manner. Popular components for an image model $\underline{b} = \left(b^{(1)}, \ldots, b^{(\underline{i})}\right)$ and $\underline{\phi} = \left(\phi^{(1)}, \ldots, \phi^{(\underline{j})}\right)$ are:

1. Data misfit terms $\mathscr{J}_{\text{data}}(\underline{b}, \underline{\phi})$
2. Terms measuring closeness to a prior model inside each subdomain $\mathscr{J}_{\text{prior}}(\underline{b}, \underline{\phi})$
3. Terms enforcing geometric constraints on the interfaces $\mathscr{J}_{\text{geom}}(\underline{b}, \underline{\phi})$

In (1), the by far most popular data misfit term is the least squares misfit cost functional which, in general, is given as an expression of the form

$$\mathscr{J}_{\text{data}}(\underline{b}, \underline{\phi}) = \frac{1}{2} \left\| \mathscr{A}(\underline{b}, \underline{\phi}) - \tilde{g} \right\|^2 = \frac{1}{2} \left\| u_M[\underline{b}, \underline{\phi}] - \tilde{g} \right\|^2, \tag{36}$$

where $\mathscr{A}(\underline{b}, \underline{\phi})$ is the forward operator defined in (2) and $u_M[\underline{b}, \underline{\phi}]$ indicates the simulated data at the set of probing locations M for this guess. Other choices can be considered as well; see, for example, [37].

(2) corresponds to classical regularization techniques, applied to each subdomain, and is treated in many textbooks, such as [37,71]. Therefore, it is not discussed in this chapter.

(3) has a long history in the shape optimization literature and in image processing applications. See, for example, [30, 87]. A few concepts are presented in Sect. 5.

Transformations and Velocity Flows

The first technique discussed here is *shape evolution by transformations and velocity flows*. This concept has been inspired by applications in continuum mechanics. Given a (possibly bounded) domain $\Omega \subset \mathbb{R}^n$ and a shape $D \subset \Omega$ with boundary ∂D which, as usual, is denoted as Γ. Let a smooth vector field $\mathbf{v} : \mathbb{R}^n \to \mathbb{R}^n$ be given with $\mathbf{v} \cdot \mathbf{n} = 0$ on $\partial\Omega$. A family of transformations S_t proceeds by

$$S_t(\mathbf{X}) = \mathbf{X} + t\mathbf{v}(\mathbf{X}) \tag{37}$$

for all $\mathbf{X} \in \Omega$. In short, $S_t = I + t\mathbf{v}$ where I stands for the identity map. This defines for each point ("particle") $\mathbf{X}$ in the domain, a propagation law prescribed by the ordinary differential equation

$$\dot{\mathbf{x}}(t, \mathbf{X}) = \mathbf{V}(t, \mathbf{x}(t, \mathbf{X})), \tag{38}$$

$$\mathbf{x}(0, \mathbf{X}) = \mathbf{X} \tag{39}$$

with the specific velocity choice

$$\mathbf{V}(t, \mathbf{x}(t, \mathbf{X})) = \mathbf{v}(\mathbf{X}). \tag{40}$$

Physically, it corresponds to the situation where each point $\mathbf{X}$ of the domain travels with constant speed along a straight line which is defined by its initial velocity vector $\mathbf{v}(\mathbf{X})$. Notice that the definition (40) can with (37) also be written in a slightly more abstract fashion as

$$\mathbf{V}(t,\mathbf{x}) = \frac{\partial}{\partial t} S_t(\mathbf{X}) = \left(\frac{\partial}{\partial t} S_t\right) \circ S^{-1}(\mathbf{x}). \tag{41}$$

In fact, it turns out that the ideas in the above example can be considerably generalized from the specific case (40) to quite arbitrary smooth vector fields $\mathbf{V}(t,\mathbf{x})$ describing smooth families of transformations $T_t(\mathbf{X})$. The generating vector field $\mathbf{V}(t,\mathbf{x})$ is often called "velocity field." It can be done as follows.

Let us given an arbitrary smooth family of transformations $T_t(\mathbf{X})$ which maps every point $\mathbf{X}$ of the domain to the point $\mathbf{x}(t,\mathbf{X}) = T_t(\mathbf{X})$ at time t. The propagation of the point $\mathbf{X}$ over time t is again described by the ordinary differential equations (38) and (39) where the velocity $\mathbf{V}$ is defined by

$$\mathbf{V}(t,\mathbf{x}) = \left(\frac{\partial}{\partial t} T_t\right) \circ T^{-1}(\mathbf{x}). \tag{42}$$

Now the propagation of points is not restricted anymore to straight lines, but can be quite arbitrary. Vice versa, given a smooth vector field $\mathbf{V}(t,\mathbf{x})$, it gives rise to a family of transformations $T_t(\mathbf{X})$ via the differential equations (38) and (39) where every point $\mathbf{X} \in \Omega$ is mapped by $T_t(\mathbf{X})$ to the solution $\mathbf{x}(t,\mathbf{X})$ of (38), (39) at time t, i.e., $T_t(\mathbf{X})(\mathbf{x}) = \mathbf{x}(t,\mathbf{X})$. For more details on this duality of transformations and velocity flows, see the well-known monographs [30, 87].

Notice that the numerical treatment of such a velocity flow in the level set framework leads to a Hamilton–Jacobi-type equation. Some remarks regarding this link are given in section "The Level Set Framework for Shape Evolution."

Eulerian Derivatives of Shape Functionals

Given the framework defined in section "Transformations and Velocity Flows," the goal is now to define transformations and velocity flows which point into a descent direction for a given cost functional. Some useful concepts on how to obtain such descent directions are discussed here.

Let $D = D_0$ be a shape embedded in the domain at time $t = 0$. When the points in the said domain start moving under the propagation laws discussed above, the interior points of the shape, the boundary points, as well as the exterior points will move as well, and therefore the shape will deform. Denote the shape at time t by $D_t = T_t(D_0)$ where as before T_t is the family of transformations which correspond to a given velocity field $\mathbf{V}(t,\mathbf{x})$. Assume furthermore that a cost functional $\mathscr{J}(\mathbf{x}, t, D_t, \ldots)$ is given which depends (among others) upon the current shape D_t. Deformation of shape will entail change of this cost. The so-called shape sensitivity analysis of structural optimization aims at quantifying these changes in the cost due to a given velocity flow (or family of transformations) in order to determine suitable descent flows.

Given a vector field $\mathbf{V}(t,\mathbf{x})$, the *Eulerian derivative* of the cost functional $\mathscr{J}(D_t)$ at time $t = 0$ in the direction $\mathbf{V}$ is defined as the limit

$$d\mathscr{J}(D,\mathbf{V}) = \lim_{t\downarrow 0} \frac{\mathscr{J}(D_t) - \mathscr{J}(D)}{t}, \tag{43}$$

if this limit exists. The functional $\mathscr{J}(D_t)$ is *shape differentiable* (or simply *differentiable*) if the Eulerian derivative $d\mathscr{J}(D,\mathbf{V})$ exists for all directions $\mathbf{V}$ and furthermore the mapping $\mathbf{V} \to d\mathscr{J}(D,\mathbf{V})$ is linear and continuous (in appropriate function spaces). It is shown in [30, 87] that if $\mathscr{J}(D)$ is shape differentiable, there exists a distribution $G(D)$ which is concentrated (supported) on $\Gamma = \partial D$ such that

$$d\mathscr{J}(D,\mathbf{V}) = \langle G(D), \mathbf{V}(0)\rangle\,. \tag{44}$$

This distribution G is the *shape gradient* of $\mathscr{J}$ in D, which is a vector distribution. More specifically, let ϖ_Γ denote the trace (or restriction) operator on the boundary Γ. Then, the Hadamard-Zolésio structure theorem states that (under certain conditions) there exists a scalar distribution g such that the shape gradient G writes in the form $G = \varpi_\Gamma^*(g\mathbf{n})$, where ϖ^* is the transpose of the trace operator at Γ and where $\mathbf{n}$ is the normal to Γ. For more details, see again [30, 87].

The Material Derivative Method

A useful concept for calculating Eulerian derivatives for cost functionals is the so-called material and shape derivative of states $\mathbf{u}$. In the application of inverse problems, these states $\mathbf{u}$ typically are the solutions of the PDEs (IEs) which model the probing fields and which depend one way or another on the shape D.

Let as before $\mathbf{V}$ be a smooth vector field with $\langle \mathbf{V}, \mathbf{n}\rangle = 0$ on $\partial\Omega$, and let $T_t(\mathbf{V})$ denote the corresponding family of transformations. Moreover, let $\mathbf{u} = \mathbf{u}[D_t]$ be a state function (of some Sobolev space) which depends on the shape $D_t \subset \Omega$ (denote as before $D_0 = D$). The *material derivative* $\dot{\mathbf{u}}[D,\mathbf{V}]$ of $\mathbf{u}$ in the direction $\mathbf{V}$ is defined as

$$\dot{\mathbf{u}}[D,\mathbf{V}] = \lim_{t\downarrow 0} \frac{\mathbf{u}[D_t]\circ T_t(\mathbf{V}) - \mathbf{u}[D]}{t}, \tag{45}$$

or

$$\dot{\mathbf{u}}[D,\mathbf{V}](\mathbf{X}) = \lim_{t\downarrow 0} \frac{\mathbf{u}[D_t](T_t(\mathbf{X})) - \mathbf{u}[D](\mathbf{X})}{t} \quad \text{for } \mathbf{X}\in\Omega\,, \tag{46}$$

where the square brackets in the notation indicate the dependence of the states and derivatives on the shape D_t and/or on the vector field $\mathbf{V}$. The material derivative corresponds to a Lagrangian point of view describing the evolution of the points in a moving coordinate system, e.g., located in the point $\mathbf{x}(t,\mathbf{X}) = T_t(\mathbf{X})$.

The *shape derivative* $\mathbf{u}'[D,\mathbf{V}]$ of $\mathbf{u}$ in the direction $\mathbf{V}$ in contrast corresponds to an Eulerian point of view observing the evolution from a fixed coordinate system, e.g., located in the point $\mathbf{X}$. It is defined as

$$\mathbf{u}'[D,\mathbf{V}] = \lim_{t\downarrow 0} \frac{\mathbf{u}[D_t] - \mathbf{u}[D]}{t}, \tag{47}$$

or

$$\mathbf{u}'[D,\mathbf{V}](\mathbf{X}) = \lim_{t\downarrow 0} \frac{\mathbf{u}[D_t](\mathbf{X}) - \mathbf{u}[D](\mathbf{X})}{t} \quad \text{for } \mathbf{X} \in \Omega. \tag{48}$$

The shape derivative and the material derivative are closely related to each other. It can be shown that

$$\mathbf{u}'[D,\mathbf{V}] = \dot{\mathbf{u}}[D,\mathbf{V}] - \nabla(\mathbf{u}[D]) \cdot \mathbf{V}(0) \tag{49}$$

provided that these quantities exist and are well defined. Subtracting $\nabla(\mathbf{u}[D]) \cdot \mathbf{V}(0)$ in (49) from the material derivative makes sure that the shape derivative actually becomes zero in the special case that the states $\mathbf{u}$ do *not* depend on the shape D. The material derivative usually does not vanish in these situations.

Some Useful Shape Functionals

To become more specific, some useful examples for shape functionals which have been applied to shape inverse problems are provided herein.

1. Define for a given function ζ the shape integral

$$\mathscr{J}_1(D) = \int_\Omega \chi_D(\mathbf{x})\zeta(\mathbf{x})d\mathbf{x} = \int_D \zeta(\mathbf{x})d\mathbf{x} \tag{50}$$

where χ_D is the characteristic function for the domain D. Then the Eulerian derivative is given by

$$d\,\mathscr{J}_1(D,\mathbf{V}) = \int_D \operatorname{div}(\zeta\mathbf{V}(0))\,d\mathbf{x} = \int_\Gamma \zeta\langle\mathbf{V}(0),\mathbf{n}\rangle_{\mathbb{R}^n} d\Gamma. \tag{51}$$

2. Consider the shape functional

$$\mathscr{J}_2(D) = \int_\Gamma \zeta(\mathbf{x})d\Gamma \tag{52}$$

for a sufficiently smooth function ζ defined on Ω such that the traces on Γ exist and are integrable. The *tangential divergence* $\operatorname{div}_\Gamma\mathbf{V}$ of the vector field $\mathbf{V}$ at the boundary Γ is defined as

$$\operatorname{div}_\Gamma\mathbf{V} = (\operatorname{div}\mathbf{V} - \langle\mathbf{DV}\cdot\mathbf{n},\,\mathbf{n}\rangle)|_\Gamma \tag{53}$$

where $\mathbf{DV}$ denotes the Jacobian of $\mathbf{V}$. Then,

$$d\mathscr{J}_2(D,\mathbf{V}) = \int_\Gamma (\langle \nabla\zeta,\, \mathbf{V}(0)\rangle + \zeta \mathrm{div}_\Gamma \mathbf{V}(0))\, d\Gamma \tag{54}$$

Be $\mathscr{N}$ an extension of the normal vector field $\mathbf{n}$ on Γ to a local neighborhood of Γ. Then, the *mean curvature* κ of Γ is defined as $\kappa = \mathrm{div}_\Gamma \mathscr{N}|_\Gamma$. With that, $d\mathscr{J}_2(D,\mathbf{V})$ admits the alternative representation

$$d\mathscr{J}_2(D,\mathbf{V}) = \int_\Gamma \left(\frac{\partial\zeta}{\partial n} + \zeta\kappa\right) \langle \mathbf{V}(0), \mathbf{n}\rangle\, d\Gamma \tag{55}$$

3. A useful link between the shape derivative and the Eulerian derivative of the cost functional is

$$\mathscr{J}_3(D) = \int_D \mathbf{u}[D] d\mathbf{x} \tag{56}$$

which depends via the states $\mathbf{u}[D]$ on the shape D. Furthermore [30, 87]

$$d\mathscr{J}_3(D,\mathbf{V}) = \int_D \mathbf{u}'[D,\mathbf{V}] d\mathbf{x} + \int_\Gamma \mathbf{u}[D]\langle \mathbf{V}(0),\, \mathbf{n}\rangle_{\mathbb{R}^n}\, d\Gamma. \tag{57}$$

4. Consider a cost functional

$$\mathscr{J}_4(D) = \int_\Gamma \zeta(\Gamma) d\Gamma \tag{58}$$

where ζ is only defined at the shape boundary Γ. Then we cannot use the characterization (49) directly, since $\nabla(\zeta)\cdot\mathbf{V}(0)$ is not well defined. In that case, the shape derivative is defined as

$$\zeta'[\Gamma,\mathbf{V}] = \dot{\zeta}[\Gamma,\mathbf{V}] - \nabla_\Gamma(\zeta[\Gamma])\cdot\mathbf{V}(0), \tag{59}$$

∇_Γ being the gradient along the boundary Γ of the shape (chosen such that $\nabla\zeta = \nabla_\Gamma\zeta + \frac{\partial\zeta}{\partial\mathbf{n}}\mathbf{n}$ whenever all these quantities are well defined). Then, the Eulerian derivative of the cost functional $\mathscr{J}_4(D)$ can be characterized as

$$d\mathscr{J}_4(D,\mathbf{V}) = \int_\Gamma \zeta'[\Gamma,\mathbf{V}] d\Gamma + \int_\Gamma \kappa\zeta\langle \mathbf{V}(0),\mathbf{n}\rangle_{\mathbb{R}^n}\, d\Gamma \tag{60}$$

where again κ denotes the mean curvature on Γ.

The Level Set Framework for Shape Evolution

So far, shape evolution has been discussed independently of its representation by a level set technique. Any of the abovementioned shape evolutions can practically be described by employing a level set representation of the shapes.

First, some convenient representations of geometric quantities in the level set framework are listed:

1. The outward normal direction [74, 85] is given by

$$\mathbf{n}(\mathbf{x}) = \frac{\nabla\phi}{|\nabla\phi|}. \tag{61}$$

2. The local curvature $\kappa(\mathbf{x})$ of ∂D, being the divergence of the normal field $\mathbf{n}(\mathbf{x})$, is

$$\kappa(\mathbf{x}) = \nabla\cdot\mathbf{n}(\mathbf{x}) = \nabla\cdot\left(\frac{\nabla\phi}{|\nabla\phi|}\right). \tag{62}$$

3. The following relation is often useful

$$\delta(\phi) = \frac{\delta_{\partial D}(\mathbf{x})}{|\nabla\phi(\mathbf{x})|} \tag{63}$$

where $\delta_{\partial D}$ is the n-dimensional Dirac delta distribution concentrated on ∂D.

Notice that the right-hand sides of (61) and (62) make sense at every point of the domain Ω where the level set function ϕ is sufficiently smooth, giving rise to a natural extension of these quantities from the boundary ∂D to a local neighborhood.

Assume now that a sufficiently smooth flow field $\mathbf{V}(\mathbf{x}, t)$ is given and that a shape D is represented by the *continuously differentiable level set function* ϕ with $|\nabla\phi| \neq 0$ at the boundary of the shape. Then, the deformation of the shape due to the flow field $\mathbf{V}(\mathbf{x}, t)$ in the level set framework can be obtained as follows.

Since the velocity fields are assumed to be sufficiently smooth, a boundary point $\mathbf{x}$ remains at the boundary of $\partial D(t)$ during the evolution of the shape. Let $\phi(\mathbf{x}, t)$ be the set of level set functions describing the shape at every time of the evolution. Differentiating $\phi(\mathbf{x}, t) = 0$ with respect to t yields

$$\frac{\partial\phi}{\partial t} + \nabla\phi\cdot\frac{d\mathbf{x}}{dt} = 0. \tag{64}$$

Identifying $\mathbf{V}(\mathbf{x}, t)$ to $\frac{d\mathbf{x}}{dt}$ and using (61), one arrives at

$$\frac{\partial\phi}{\partial t} + |\nabla\phi|\,\mathbf{V}(\mathbf{x}, t)\cdot\mathbf{n}(\mathbf{x}, t) = 0. \tag{65}$$

Defining the normal velocity as

$$F(\mathbf{x}, t) = \mathbf{V}(\mathbf{x}, t)\cdot\mathbf{n}(\mathbf{x}, t) \tag{66}$$

the *Hamilton–Jacobi-type equation* for describing the evolution of the level set function follows as

$$\frac{\partial \phi}{\partial t} + F(\mathbf{x}, t) \cdot |\nabla \phi| = 0. \tag{67}$$

5 Shape Evolution Driven by Geometric Constraints

It is possible to define a shape evolution without any data misfit functional being involved. This type of shape evolution often occurs in applications of image processing or computational physics. For example, starting from an initial shape, the goal might be to define a shape evolution which aims at reducing the cost of the image with respect to one or more geometric quantities, typically encoded in some geometric cost functional. Based on the theory developed in Sect. 4, some useful expressions will be derived here for calculating such descent directions. The then obtained geometrically driven shape evolutions can also be used for adding additional constraints or regularization during the shape evolution driven by data misfit, if desired. This is achieved practically by adding appropriate geometrical cost terms to the data misfit term and calculating descent directions for this combined cost.

Penalizing Total Length of Boundaries

Assume that $\Gamma = \partial D$ is a smooth submanifold in Ω. The total length (or surface) of Γ is defined as

$$\mathscr{J}_{\mathrm{len}\Gamma}(D) = \int_{\Gamma} d\,\Gamma = \int_{\Omega} \delta_{\partial D}(\mathbf{x})\, d\,\mathbf{x}. \tag{68}$$

Applying a flow by a smooth vector field $\mathbf{V}(\mathbf{x}, t)$, Eq. (55) yields with $\zeta = 1$ an expression for the corresponding change in the cost (68) which is

$$d\,\mathscr{J}_{\mathrm{len}\Gamma}(D, \mathbf{V}) = \int_{\Gamma} \kappa\, \langle \mathbf{V}(0), \mathbf{n} \rangle\, d\,\Gamma. \tag{69}$$

If the shape D is represented by a continuously differentiable level set function ϕ, an alternative derivation can be given. First, using (63), write (68) in the form

$$\mathscr{J}_{\mathrm{len}\Gamma}(D(\phi)) = \int_{\Omega} \delta(\phi) |\nabla \phi(\mathbf{x})|\, d\,\mathbf{x}. \tag{70}$$

Perturbing now $\phi \to \phi + \psi$, formal calculation (see, e.g., [92]) yields that the cost functional is perturbed by

$$\left\langle \frac{\partial \mathscr{J}_{\text{len}\Gamma}}{\partial \phi}, \psi \right\rangle = \int_\Omega \delta(\phi)\psi(\mathbf{x})\nabla \cdot \frac{\nabla \phi}{|\nabla \phi|}\, d\mathbf{x}. \tag{71}$$

Therefore, using (62), it can be identified

$$\frac{\partial \mathscr{J}_{\text{len}\Gamma}}{\partial \phi} = \delta(\phi)\nabla \cdot \frac{\nabla \phi}{|\nabla \phi|} = \delta(\phi)\kappa \tag{72}$$

where κ is now an extension (defined, e.g., by (62)) of the local curvature to a small neighborhood of Γ. For both representations (69) and (72), minimizing the cost by a gradient method leads to curvature-driven flow equations, which is $\mathbf{V}(0) = -\kappa \mathbf{n}$. This curvature-dependent velocity has been widely used to regularize the computation of motion of fronts via the level set method [48], as well in the field of image processing [70], and has been introduced also recently for regularizing inverse problems; see, e.g., [41, 79, 82].

Two popular concepts related to the above shape evolution are the Mumford–Shah and the total variation functionals, which are frequently employed in image segmentation applications. This relationship is briefly described in the following.

The popular *Mumford–Shah functional for image segmentation* [70] contains, in addition to a fidelity term inside each region of the segmented image, a term which encourages to shorten total curve length of the interfaces. This latter term can be written for piecewise constant binary media (see section "The Basic Level Set Formulation for Binary Media" with constant profiles in each region) as

$$\mathscr{J}_{\text{MS}} = \int_\Omega |\nabla H(\phi)|\, d\mathbf{x}. \tag{73}$$

Taking into account that $\nabla H(\phi) = H'(\phi)\nabla\phi = \delta(\phi)|\nabla\phi|\mathbf{n}$, it is seen that $\mathscr{J}_{\text{MS}} = \mathscr{J}_{\text{len}\Gamma}(D(\phi))$ as given in (70), which again yields the curvature-driven flow Eq. (72). For more details, see [26, 38, 95].

The *total variation (TV) functional*, on the other hand, can be written, again for the situation of piecewise constant binary media, as

$$\mathscr{J}_{\text{TV}} = \int_\Omega |\nabla b(\phi)|\, d\mathbf{x} = |b_e - b_i| \int_\Omega |\nabla H(\phi)|\, d\mathbf{x}. \tag{74}$$

Therefore, it coincides with the Mumford–Shah functional $\mathscr{J}_{\text{MS}}$ up to the factor $|b_e - b_i|$. Roughly it can be said that the TV functional (74) penalizes the product of the jump between different regions and the arc length of their interfaces, whereas the Mumford–Shah functional (73) penalizes only this arc length. Refer for more information to [25, 38].

Penalizing Volume or Area of Shape

It is again assumed that $\Gamma = \partial D$ is a smooth submanifold in Ω. Define the total area (volume) of D as

$$\mathscr{J}_{\mathrm{vol}D}(D) = \int_D d\mathbf{x} = \int_\Omega \chi_D(\mathbf{x})\, d\mathbf{x}, \tag{75}$$

where the *characteristic function* $\chi_D : \Omega \to \{0, 1\}$ for a given shape D is defined as

$$\chi_D(\mathbf{x}) = \begin{cases} 1, & \mathbf{x} \in D \\ 0, & \mathbf{x} \in \Omega \backslash D. \end{cases} \tag{76}$$

Applying a flow by a smooth vector field $\mathbf{V}(\mathbf{x}, t)$, Eqs. (50) and (51) yield with $\zeta = 1$

$$d\,\mathscr{J}_{\mathrm{vol}D}(D, \mathbf{V}) = \int_D \mathrm{div}\mathbf{V}(0) d\mathbf{x} = \int_\Gamma \langle \mathbf{V}(0), \mathbf{n} \rangle\, d\Gamma. \tag{77}$$

Again, if the shape D is represented by a continuously differentiable level set function ϕ, an alternative derivation can be given. First, using the Heaviside function H, let us write (75) in the form

$$\mathscr{J}_{\mathrm{vol}D}(D) = \int_\Omega H(\phi)\, d\mathbf{x}. \tag{78}$$

Perturbing as before $\phi \to \phi + \psi$, it follows

$$\left\langle \frac{\partial \mathscr{J}_{\mathrm{vol}D}}{\partial \phi}, \psi \right\rangle = \int_\Omega \delta(\phi)\psi(\mathbf{x})\, d\mathbf{x} \tag{79}$$

such that

$$\frac{\partial \mathscr{J}_{\mathrm{vol}D}}{\partial \phi} = \delta(\phi). \tag{80}$$

In both formulations, a descent flow is given by a motion with constant speed in the negative direction of the normal $\mathbf{n}$ to the boundary Γ, which is $\mathbf{V}(0) = -\mathbf{n}$.

6 Shape Evolution Driven by Data Misfit

An essential goal in the solution of inverse problems is to find an image which is able to reproduce the measured data in a certain sense. As far as interfaces are concerned, this gives rise to the need of finding descent directions for shape evolution with

respect to the data misfit functional. In the following, some concepts are presented which aim at providing these descent directions during the shape evolution. These concepts can be combined arbitrarily with the above-discussed concepts for shape evolution driven by geometric terms.

Shape Deformation by Calculus of Variations

Historically, the first approach for applying a level set technique for solving an inverse problem in [82] has used concepts from the calculus of variations for calculating descent directions for the data misfit functional. In many applications, this approach is still a very convenient way of deriving evolution laws for shapes. In the following, the main ideas of this approach are briefly reviewed, following [82]. The goal is to obtain expressions for the deformation of already existing shapes according to a normal velocity field defined at the boundary of these shapes. Topological changes are not formally included in the consideration at this stage (even though they occur automatically when implementing the discussed schemes in a level set-based numerical framework). The formal treatment of topological changes is a topic of active current research and will be discussed briefly in section "Topological Derivatives."

Least Squares Cost Functionals and Gradient Directions

Typically, appropriate function spaces are needed for defining and calculating appropriate descent directions with respect to the data misfit cost functional. Without being very specific, in the following, the general notation P is used for denoting the space of parameters b and, if not otherwise specified, Z for denoting the space of measurements $\tilde{g}$. For simplicity, these function spaces are considered being appropriately chosen Hilbert or vector spaces. Certainly, other types of spaces can be used as well, which might lead to interesting variants of the described concepts.

Consider now the least squares cost functional

$$\mathscr{J}(b) = \frac{1}{2} \|\mathscr{R}(b)\|_Z^2 = \frac{1}{2} \langle \mathscr{R}(b), \mathscr{R}(b) \rangle_Z , \tag{81}$$

where $\langle , \rangle_Z$ denotes the canonical inner product in data space Z. Assume that $\mathscr{R}(b)$ admits the expansion

$$\mathscr{R}(b + \delta b) = \mathscr{R}(b) + \mathscr{R}'(b)\delta b + O\left(\|\delta b\|_P^2\right), \tag{82}$$

letting $\| \ \|_P$ be the canonical norm in parameter space P, for a sufficiently small perturbation (variation) $\delta b \in P$. The linear operator $\mathscr{R}'(b)$ (if it exists) is often called the *Fréchet derivative* of $\mathscr{R}$. Plugging (82) into (81) yields the relationship

$$\mathscr{J}(b + \delta b) = \mathscr{J}(b) + \ \mathtt{Re}\left\langle \mathscr{R}'(b)^* \mathscr{R}(b), \delta b \right\rangle_P + O\left(\|\delta b\|_P^2\right) \tag{83}$$

where the symbol Re indicates the real part of the corresponding quantity. The operator $\mathscr{R}'(b)^*$ is the formal adjoint operator of $\mathscr{R}'(b)$ with respect to spaces Z and P:

$$\left\langle \mathscr{R}'(b)^* g,\, \hat{b} \right\rangle_P = \left\langle g,\, \mathscr{R}'(b)\hat{b} \right\rangle_Z \quad \text{for all } \hat{b} \in P,\ g \in Z. \tag{84}$$

The quantity

$$\mathbf{grad}_b\, \mathscr{J} = \mathscr{R}'(b)^* \mathscr{R}(b) \tag{85}$$

is called the *gradient direction* of $\mathscr{J}$ in b.

It is assumed that the operators $\mathscr{R}'(b)$ and $\mathscr{R}'(b)^*$ take into account the correct interface conditions at ∂D, which is important when actually evaluating these derivatives in a "direct" or in an "adjoint" fashion. In many practical applications, the situation can occur that (formally) the fields need to be evaluated at interfaces where jumps occur. In these situations, appropriate limits can be considered. Alternatively, the tools developed in section "Shape Sensitivity Analysis and the Speed Method" can be applied there. The existence and special form of Fréchet derivatives $\mathscr{R}'(b)$ (and the corresponding shape derivatives) for parameter distributions b with discontinuities along interfaces are problem specific and beyond the scope of this chapter. Refer to the cited literature, for example, [9, 15, 49, 53, 54, 58, 78]. In many practical implementations, the interface ∂D is de facto replaced by a narrow transition zone with smoothly varying parameters, in which case the interface conditions disappear.

Change of b Due to Shape Deformations

Assume that every point $\mathbf{x}$ moves in the domain Ω a small distance $\mathbf{y}(\mathbf{x})$ and that the mapping $\mathbf{x} \to \mathbf{y}(\mathbf{x})$ is sufficiently smooth, such that the basic structure of the shape D remains preserved. Then, the points located on the boundary $\Gamma = \partial D$ will move to the new locations $\mathbf{x}' = \mathbf{x} + \mathbf{y}(\mathbf{x})$, and the boundary Γ will be deformed into the new boundary $\Gamma' = \partial D'$. Assume furthermore that the parameter distribution in Ω has the special form (4), such that it will change as well. In the following, the first goal is to quantify this change in the parameter distribution $b(\mathbf{x})$ due to an infinitesimal deformation as described above.

Consider the inner product of δb with a test function f

$$\langle \delta b,\, f \rangle_\Omega = \int_\Omega \delta b(\mathbf{x})\, \overline{f(x)}\, d\mathbf{x} = \int_{\mathrm{symdiff}(D,D')} \delta b(\mathbf{x})\, \overline{f(\mathbf{x})}\, d\mathbf{x}, \tag{86}$$

where the overline means "complex conjugate" and $\mathrm{symdiff}(D, D') = (D \cup D')\backslash(D \cap D')$ is the symmetric difference of the sets D and D' (see Fig. 8). Since the difference in D and D' is infinitesimal, the area integral reduces to a line integral. Let $\mathbf{n}(\mathbf{x})$ denote the outward normal to $\mathbf{x}$. Then, the integral in (86) becomes

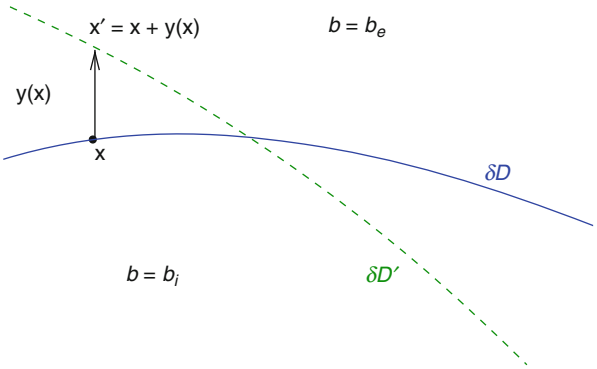


Fig. 8 Deformation of shapes using calculation of small variations

$$\langle \delta b,\ f \rangle_{\partial D} = \int_{\delta D} \Big(b_i(\mathbf{x}) - b_e(\mathbf{x}) \Big) \mathbf{y}(\mathbf{x}) \cdot \mathbf{n}(\mathbf{x}) \overline{f(\mathbf{x})}\, ds(\mathbf{x}), \tag{87}$$

where $ds(\mathbf{x})$ is the incremental arclength. Here it has been used that in the limit $\delta b(\mathbf{x}) = b_i(\mathbf{x}) - b_e(\mathbf{x})$ at the boundary point $\mathbf{x} \in \partial D$ due to (4). It follows the result

$$\delta b(\mathbf{x}) = \varpi_{\partial D} \Big(\big(b_i(\mathbf{x}) - b_e(\mathbf{x}) \big)\, \mathbf{y}(\mathbf{x}) \cdot \mathbf{n}(\mathbf{x}) \Big) \tag{88}$$

where $\varpi_{\partial D}$ is the n-dimensional restriction operator which restricts functions defined in Ω to the boundary ∂D of the shape D ($n = 2$ or 3, usually). Therefore, $\delta b(\mathbf{x})$ is interpreted now as a surface measure on ∂D. Using the n-dimensional Dirac delta distribution $\delta_{\partial D}$ concentrated on the boundary ∂D of the shape D, (88) can be written in the form

$$\delta b(\mathbf{x}) = (b_i - b_e)\, \mathbf{y}(\mathbf{x}) \cdot \mathbf{n}(\mathbf{x})\, \delta_{\partial D}(\mathbf{x}) \tag{89}$$

which is a distribution defined on the entire domain Ω but concentrated on ∂D where it has the same strength as the corresponding surface measure. Although, strictly speaking, they are different mathematical objects, they are identified in the following for simplicity. Compare (87) also to the classical shape or domain derivative as, for example, calculated in [49], focusing there on the effect of the infinitesimal change in the boundary of a scatterer on the far-field pattern of a scattering experiment.

Variation of Cost Due to Velocity Field v(x)

A popular approach for generating small displacements $\mathbf{y}(\mathbf{x})$ (as discussed in section "Change of b Due to Shape Deformations") for moving the boundary ∂D is to assign to each point in the domain a *velocity field* $\mathbf{v}(\mathbf{x})$ and to let the points $\mathbf{x} \in \Omega$ move a small artificial evolution time $[0, \tau]$ with constant velocity $\mathbf{v}(\mathbf{x})$. Then

$$\mathbf{y}(\mathbf{x}) = \mathbf{v}(\mathbf{x})\tau. \tag{90}$$

Plugging this into (89) for $t \in [0, \tau]$, the corresponding change in the parameters follows as

$$\delta b(\mathbf{x}; t) = (b_i - b_e)\, \mathbf{v}(\mathbf{x}) \cdot \mathbf{n}(\mathbf{x}) t\, \delta_{\partial D}(\mathbf{x}). \tag{91}$$

Plugging expression (91) into (83) and neglecting terms of higher than linear order yields

$$\begin{aligned} \mathscr{J}(b(t)) - \mathscr{J}(b(0)) &= \mathtt{Re}\left\langle \mathbf{grad}_b \mathscr{J}\,,\, \delta b(\mathbf{x}; t) \right\rangle_P \\ &= \mathtt{Re}\left\langle \mathbf{grad}_b \mathscr{J}\,,\, (b_i - b_e)\, \mathbf{v}(\mathbf{x}) \cdot \mathbf{n}(\mathbf{x}) t\, \delta_{\partial D}(\mathbf{x}) \right\rangle_P \end{aligned} \tag{92}$$

or, in the limit $t \to 0$, evaluating the Dirac delta distribution,

$$\left.\frac{\partial \mathscr{J}(b)}{\partial t}\right|_{t=0} = \mathtt{Re} \int_{\partial D} \mathbf{grad}_b \mathscr{J}\, \overline{(b_i - b_e)} \mathbf{v}(\mathbf{x}) \cdot \mathbf{n}(\mathbf{x}) ds(\mathbf{x}), \tag{93}$$

where the overline means "complex conjugate" and $\mathbf{grad}_b \mathscr{J}$ is defined in (85). Similar expressions will be derived further below using formal shape sensitivity analysis. (Compare, e.g., for the situation of TM-waves, the expression (97) calculated by using (93) with the analogous expressions (100) and (105) calculated by using formal shape sensitivity analysis.)

If a velocity field $\mathbf{v}(\mathbf{x})$ can be found such that $\left.\frac{\partial \mathscr{J}(b)}{\partial t}\right|_{t=0} < 0$, then it is expected (for continuity reasons) that this inequality holds in a sufficiently small time interval $[0, \tau]$ and that therefore the total cost during the artificial flow will be reduced. This will be the general strategy in most optimization type approaches for solving the underlying inverse problem. See the brief discussion in section "Shape Evolution and Shape Optimization."

Notice that only the normal component of the velocity field

$$F(\mathbf{x}) = \mathbf{v}(\mathbf{x}) \cdot \mathbf{n}(\mathbf{x}) \tag{94}$$

at the boundary ∂D of the shape D is of relevance for the change in the cost (compare the remarks made already in section "The Level Set Framework for Shape Evolution"). This is because tangential components of $\mathbf{v}$ do not contribute to shape deformations. In a parameterized way of thinking, they only "re-parameterize" the existing boundary.

Example: Shape Variation for TM-Waves

An instructive example is given here in order to demonstrate the above concepts for a practical application. Consider TM-waves in a typical imaging situation of subsurface imaging or of microwave breast screening as in the case study of section "Example 1: Microwave Breast Screening." Assume for simplicity that the basic level set model of section "The Basic Level Set Formulation for Binary Media"

is applied here. The cost functional measuring the mismatch between calculated data $u_M[D]$ corresponding to the shape D and physically measured data $\tilde{g}$ is defined as

$$\mathscr{J}(D) = \frac{1}{2} \|u_M[D] - \tilde{g}\|^2_{L^2(M)}, \tag{95}$$

where the calculated measurements u_M are given as the electric field values $u(\mathbf{x})$ at the set of receiver locations M. Using a properly defined adjoint state $z(\mathbf{x})$ (see, e.g., [30, 34, 71] for details on adjoint states), it can be shown by straightforward calculation that $\mathscr{R}'(b)^* \mathscr{R}(b)$ takes the form

$$\left(\mathbf{grad}_b \mathscr{J}\right)(\mathbf{x}) = \overline{u(\mathbf{x}) z(\mathbf{x})}, \tag{96}$$

where $u(\mathbf{x})$ denotes the solution of the forward problem and $\mathbf{grad}_b \mathscr{J}$ is defined as in (85). Therefore, it follows that

$$\left.\frac{\partial \mathscr{J}(b)}{\partial t}\right|_{t=0} = \mathtt{Re} \int_{\partial D} u(\mathbf{x}) z(\mathbf{x}) (b_i - b_e) \mathbf{v}(\mathbf{x}) \cdot \mathbf{n}(\mathbf{x}) ds(\mathbf{x}), \tag{97}$$

where it is used that the real part of a complex number and its complex conjugate are identical. Similar expressions based on adjoint field calculations can be derived for a large variety of applications; see, for example, [34, 39, 45, 71, 84, 89, 90].

Example: Evolution of Thin Shapes (Cracks)

Another application of this technique has been presented in [4, 77] for finding cracks in a homogeneous material from boundary data; see section "Example 3: Crack Detection." The evolution of cracks as defined in section "A Modification of the Classical Level Set Technique for Describing Cracks or Thin Shapes" requires the simultaneous consideration of two level set functions. Evolution of the first level set function amounts to displacement of the thin region (crack) in the transversal direction, whereas evolution of the second level set function describes the process of crack growth or shrinkage in the longitudinal direction, which comes with the option of crack splitting and merging. Descent directions for both level set functions can be calculated by the following arguments presented above. It needs to be taken into account, however, that, due to the specific construction of a crack with finite thickness, deformation of the zero level set of the first level set function is associated with a displacement of the crack boundary at two adjacent locations, which both contribute to a small variation in the cost. See Fig. 9.

Assume that a small displacement is applied to the zero level set of the first level set function defining S in the notation of section "A Modification of the Classical Level Set Technique for Describing Cracks or Thin Shapes." This is reflected by two contributions to the least squares data misfit cost, one from the deformation of S^- in Fig. 9 and the other one from the deformation of S^+ in Fig. 9. It follows that a descent velocity is now given as $\mathbf{v}(\mathbf{x}) = F_{\varphi_1}(\mathbf{x})\mathbf{n}(\mathbf{x})$ with

$$F_{\varphi_1}(\mathbf{x}) = -(b_i - b_e)\left[\mathbf{grad}_b \mathscr{J}|_{S^+} - \mathbf{grad}_b \mathscr{J}|_{S^-}\right] \quad \text{on } S, \tag{98}$$

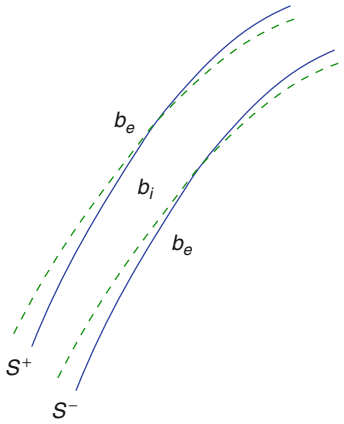


Fig. 9 Deformation of a thin shape (crack) using calculus of small variations

with $\mathbf{grad}_b \mathscr{J}$ being defined in (85). In (98), for each $\mathbf{x} \in \Gamma_{\phi_1}$, two adjacent points of $\mathbf{grad}_b \mathscr{J}|_{S^+}$ and $\mathbf{grad}_b \mathscr{J}|_{S^-}$ contribute to the value of $F_{\varphi_1}(\mathbf{x})$ which can be found in the normal direction to Γ_{ϕ_1} in $\mathbf{x}$.

In a similar way, a descent direction with respect to the second level set function ϕ_2 can be obtained. Its detailed derivation depends slightly on the way how the crack tips are constructed in section "A Modification of the Classical Level Set Technique for Describing Cracks or Thin Shapes," where two alternative choices are given. Overall, a descent velocity can be calculated following classical rules for those points $\mathbf{x}$ of ∂B (i.e., for those points of the zero level set of ϕ_2) which satisfy $\mathbf{x} \in \partial B \cap \Gamma^{\delta}_{\phi_1}$ or, alternatively, $\mathbf{x} \in \partial B \cap \Gamma_{\phi_1}$. Then, the obtained velocity field needs to be extended first to the remaining parts of ∂B and then to the rest of Ω. Notice that the specific form of $\mathbf{grad}_b \mathscr{J}$ might be slightly different here from the one given in (96) due to the slightly different PDE which might be involved here (depending on the application). For more details, refer to [4, 77].

Shape Sensitivity Analysis and the Speed Method

In this section, an alternative technique is presented for formally defining shape derivatives and modeling shape deformations driven by cost functionals. This theory, called *shape sensitivity analysis*, is quite general and powerful, such that it is used heavily in various applications. Only very few concepts of it can be mentioned here which are employed when calculating descent directions with respect to data misfit cost functionals.

The tools, as presented here, have been used and advanced significantly during the last twenty years in the framework of optimal shape design [30, 87]. Having this powerful theory readily available, it is therefore quite natural that these methods have been applied very early already to the applications of shape-based inverse problems with level sets. The theory of this section is again mainly concentrated

upon modeling in a formally accurate way the deformation of already existing shapes. It does not incorporate topological changes. These will be discussed briefly in section "Topological Derivatives."

Example: Shape Sensitivity Analysis for TM-Waves

Again the situation of inverse scattering by TM-waves is considered here. The below discussion closely follows the results presented in [64]. The main tools used here are the material and shape derivative defined in section "The Material Derivative Method."

The cost functional measuring the mismatch between calculated data $u_M[D]$ corresponding to the shape D and physically measured data $\tilde{g}$ is defined by (95). When perturbing the shape by a velocity field $\mathbf{V}(t, \mathbf{x})$, the electric field at the (fixed) probing line changes according to $u \to u + u'$, where u' is the shape derivative defined in section "The Material Derivative Method." Plugging this into (95) and neglecting terms of higher than linear order, it is verified that

$$d\mathscr{J}(D, \mathbf{V}) = \quad \mathtt{Re} \int_M u'(\mathbf{x}) \, \overline{(u_M - \tilde{g})}(\mathbf{x}) d\mathbf{x}. \tag{99}$$

Now, the shape derivative u' can be calculated by first computing the material derivative (also defined in section "The Material Derivative Method") and then using one of the relationships between the material derivative and the shape derivative (see sections "The Material Derivative Method" and "Some Useful Shape Functionals"). Using also here an adjoint state z, the Eulerian derivative can be characterized and calculated as

$$d\mathscr{J}(D, \mathbf{V}) = \quad \mathtt{Re} \int_\Gamma (b_{\text{int}} - b_{\text{ext}}) u(\mathbf{x}) z(\mathbf{x}) \mathbf{V}(0, \mathbf{x}) \cdot \mathbf{n} \, d\Gamma. \tag{100}$$

Notice that this is exactly the same result as we arrived at in (97). For more details, refer to [64].

Shape Derivatives by a Min–Max Principle

In order to avoid the explicit calculation of material and shape derivatives of the states with respect to the flow fields, an alternative technique can be used as reported in [29, 30, 78]. It is based on a reformulation of the derivative of a shape functional $\mathscr{J}(D)$ with respect to time as the partial derivative of a saddle point (or a "min–max") of a suitably defined Lagrangian. In the following, the basic features of this approach will be outlined, focusing in particular on the shape derivative for TM-waves.

Let again the cost functional $\mathscr{J}(D(t))$ be defined as in (95) by

$$\mathscr{J}(D(t)) = \frac{1}{2} \left\| u_M[D(t)] - \tilde{g} \right\|_{L^2(M)}. \tag{101}$$

The goal is to write $\mathscr{J}(D(t))$ in the form

$$\mathscr{J}(D(t)) = \min_u \max_z \mathscr{L}(t,u,z) \tag{102}$$

for some suitably defined Lagrangian $\mathscr{L}(t,u,z)$. Here and in the following, the complex nature of the forward fields u and the adjoint fields z is (partly) neglected in order to simplify notation (more rigorous expressions can be found in [78]). The Lagrangian $\mathscr{L}(t,u,z)$ takes the form

$$\mathscr{L}(t,u,z) = \frac{1}{2}\int_M \left| \int_\Omega \eta\left(\mathbf{x}'\right) G_{12}\left(\mathbf{x},\mathbf{x}'\right) u\left(\mathbf{x}'\right)\, d\mathbf{x}' - \tilde{g}(x) \right|^2 dx \tag{103}$$

$$+ \ \mathrm{Re}\int_\Omega \left(u(\mathbf{x}) - u^{\mathrm{inc}}(\mathbf{x}) - \int_\Omega \eta\left(\mathbf{x}'\right) G_{22}\left(\mathbf{x},\mathbf{x}'\right) u\left(\mathbf{x}'\right) d\mathbf{x}' \right) z(\mathbf{x})\, dx.$$

Next, it can be shown that this Lagrangian has a unique saddle point denoted by (u^*, z^*), which is characterized by an optimality condition with respect to u and z. In fact, the uniqueness follows from the well-posedness and uniqueness of the solutions of the direct and adjoint state equations; see [31, 78]. The key observation is now that

$$\frac{d\mathscr{J}}{dt} = \frac{\partial}{\partial t}\left(\min_u \max_z \mathscr{L}(t,u,z) \right) = \frac{\partial}{\partial t}\mathscr{L}(t,u^*,z^*) \tag{104}$$

which says that the time derivative of the original cost functional can be replaced by the partial time derivative of a saddle point. Following these ideas, the result is derived

$$\frac{d\mathscr{J}}{dt} = \ \mathrm{Re}\int_\Gamma (b_{\mathrm{int}} - b_{\mathrm{ext}}) u(\mathbf{x}) z(\mathbf{x}) \mathbf{V}(0)\cdot\mathbf{n}\, d\Gamma \tag{105}$$

which holds for *TM-waves* and which is identical to the previously derived expressions (97) and (100).

Similar expressions (now involving expressions of the form $\nabla u(\mathbf{x})\nabla z(\mathbf{x})$ rather than $u(\mathbf{x})z(\mathbf{x})$ at the interfaces) can be derived also for so-called TE-waves; see [78]. The above outlined min–max approach is in fact wide ranging and can be extended to 3D vector scattering, geometrical regularizations, simultaneous searches of shape and contrast, etc. It de facto applies as soon as one has well-posedness of the direct and adjoint problems. For more details, refer to [29, 30, 78, 79].

Formal Shape Evolution Using the Heaviside Function

A third possibility for describing and modeling shape deformations driven by data misfit (in addition to using calculus of variation as in section "Shape Deformation by

Calculus of Variations" or shape sensitivity analysis as in section "Shape Sensitivity Analysis and the Speed Method") is the use of the characteristic function and formal (basic) distribution theory. In contrast to the previous two techniques which first calculate velocity fields in the normal direction to the interfaces and then move the interfaces accordingly using a level set technique (or any other computational front propagation technique), typically leading to a Hamilton–Jacobi-type formalism (compare the remarks in section "The Level Set Framework for Shape Evolution"), the method presented in the following does not explicitly use the concept of velocity vector fields, but instead tries to design evolution laws directly for the describing level set functions (thereby not necessarily leading to Hamilton–Jacobi-type evolution laws).

Notice that many of the level set formulations presented in Sect. 3 give rise to similar concepts as discussed in the following. On the other hand, typically also the concepts discussed in sections "Shape Deformation by Calculus of Variations" and "Shape Sensitivity Analysis and the Speed Method" can be translated, once suitable velocity fields have been determined, into level set evolutions using the various representations of Sect. 3. Details on how these evolution laws can be established can be found in the literature cited in Sect. 3.

The formalism discussed in the following is in fact very flexible and quite easy to handle if standard rules for calculations with distributions are taken into account. Moreover, it often leads to very robust and powerful reconstruction algorithms. Certainly, it also has some limitations: in the form presented here, it is mainly applicable for "penetrable objects" with finite jumps in the coefficients between different regions. This means that it does not generally handle inverse scattering from impenetrable obstacles if very specific and maybe quite complicated boundary conditions need to be taken into account at the scatterer surfaces. For those applications, the theory based on shape sensitivity analysis is more appropriate. Nevertheless, since many inverse scattering problems can be described in the form presented in the following, possibly incorporating some finite jump conditions of the forward and adjoint fields or their normal components across the interfaces (which can be handled by slightly "smearing out" these interfaces over a small transition zone), this theory based on formal distribution theory provides an interesting alternative when deriving level set-based shape evolution equations for solving inverse scattering problems.

The main idea of this technique is demonstrated by giving two examples related to the test cases from sections "Example 1: Microwave Breast Screening" and "Example 2: History Matching in Petroleum Engineering."

Example: Breast Screening–Smoothly Varying Internal Profiles

In the example of breast screening as discussed in section "Example 1: Microwave Breast Screening," three level set functions and four different interior parameter profiles need to be reconstructed from the given data simultaneously. Some of the interior profiles are assumed to be constant, whereas others are smoothly varying. In the following, the theory is developed under the assumption that all interior parameter profiles are smoothly varying. The case of constant parameter profiles

in some region is captured in the next section where a similar case is discussed for the application of reservoir engineering.

Let $\mathscr{R}(b(\phi_1, \phi_2, \phi_3, b_1, b_2, b_3, b_4))$ denote the difference between measured data and data corresponding to the latest best guess $(\phi_1, \phi_2, \phi_3, b_1, b_2, b_3, b_4)$ of level set functions and interior profiles in the image model discussed in section "A Modification of Color Level Set for Tumor Detection." Then, the least squares data misfit for tumor detection is given by

$$\mathscr{J}(b(\phi_1, \phi_2, \phi_3, b_1, b_2, b_3, b_4)) = \frac{1}{2} \|\mathscr{R}(b(\phi_1, \phi_2, \phi_3, b_1, b_2, b_3, b_4))\|^2 . \tag{106}$$

Introducing an artificial evolution time t for the above specified unknowns of the inverse problem, the goal is to find evolution laws

$$\frac{d\phi_\nu}{dt} = f_\nu(\mathbf{x}, t), \ \nu = 1, \ldots, 3, \tag{107}$$

$$\frac{db_\nu}{dt} = g_\nu(\mathbf{x}, t), \ \nu = 1, \ldots, 4, \tag{108}$$

such that the cost $\mathscr{J}$ decreases with increasing evolution time. With level set functions and interior profiles evolving, also the cost will change, $\mathscr{J} = \mathscr{J}(t)$, such that formally its time derivative can be calculated by using the chain rule

$$\begin{aligned} \frac{d\mathscr{J}}{dt} &= \frac{d\mathscr{J}}{db} \left[\sum_{\nu=1}^{3} \frac{\partial b}{\partial \phi_\nu} \frac{d\phi_\nu}{dt} + \sum_{\nu=1}^{4} \frac{\partial b}{\partial b_\nu} \frac{db_\nu}{dt} \right] \\ &= \mathtt{Re} \left\langle \mathbf{grad}_b \mathscr{J}, \sum_{\nu=1}^{3} \frac{\partial b}{\partial \phi_\nu} f_\nu + \sum_{\nu=1}^{4} \frac{\partial b}{\partial b_\nu} g_\nu \right\rangle_P . \end{aligned} \tag{109}$$

Here, Re indicates to take the real part of the following complex quantity and $\langle \ , \ \rangle_P$ denotes a suitable inner product in parameter space P, and $\mathbf{grad}_b \mathscr{J}$ is defined in (85). It is verified easily that in the situation of section "A Modification of Color Level Set for Tumor Detection"

$$\begin{aligned} \frac{\partial b}{\partial \phi_1} &= \delta(\phi_1) \Big(-b_1 + b_2(1 - H(\phi_2)) \\ &\quad + H(\phi_2) \Big\{ b_3(1 - H(\phi_3)) + b_4 H(\phi_3) \Big\} \Big), \end{aligned} \tag{110}$$

$$\frac{\partial b}{\partial \phi_2} = H(\phi_1)\delta(\phi_2) \left[-b_2 + b_3, (1 - H(\phi_3)) + b_4 H(\phi_3) \right], \tag{111}$$

$$\frac{\partial b}{\partial \phi_3} = H(\phi_1) H(\phi_2) \delta(\phi_3) \left\{ -b_3 + b_4 \right\}, \tag{112}$$

and

$$\frac{\partial b}{\partial b_1} = 1 - H(\phi_1), \tag{113}$$

$$\frac{\partial b}{\partial b_2} = H(\phi_1)(1 - H(\phi_2)), \tag{114}$$

$$\frac{\partial b}{\partial b_3} = H(\phi_1)H(\phi_2)(1 - H(\phi_3)), \tag{115}$$

$$\frac{\partial b}{\partial b_4} = H(\phi_1)H(\phi_2)H(\phi_3). \tag{116}$$

Descent directions are therefore given by

$$f_\nu(t) = -C_\nu(t) \ \ \mathtt{Re}\left[\mathbf{grad}_b \mathscr{J} \frac{\partial b}{\partial \phi_\nu}\right], \ \nu = 1, \ldots, 3, \tag{117}$$

$$g_\nu(t) = -\hat{C}_\nu(t) \ \ \mathtt{Re}\left[\mathbf{grad}_b \mathscr{J} \frac{\partial b}{\partial b_\nu}\right], \ \nu = 1, \ldots, 4, \tag{118}$$

with some appropriately chosen positive-valued speed factors $C_\nu(t)$ and $\hat{C}_\nu(t)$. An efficient way to compute $\mathbf{grad}_b \mathscr{J}$ is again to use the *adjoint formulation*; see (96) and for more details [34, 52, 71].

Notice that is might be convenient to approximate the Dirac delta on the right-hand side of (110)–(112) in the formulation of the level set evolution by either a norrowband scheme or by a positive constant which allows for topological changes in the entire computational domain driven by the least squares data misfit. For more details, see the brief discussion in section "Shape Evolution and Shape Optimization" and the slightly more detailed discussions held in [31,52]. Following the latter scheme, one possible numerical discretization of the expressions (107) and (108) in time $t = t^{(n)}$, $n = 0, 1, 2, \ldots$, then yields the update rules

$$\phi_\nu^{(n+1)} = \phi_\nu^{(n)} - \delta t^{(n)} C_\nu\left(t^{(n)}\right) \ \ \mathtt{Re}\left[\mathbf{grad}_b \mathscr{J} \frac{\partial b}{\partial \phi_\nu}\right]^{(n)}, \ \nu = 1, \ldots, 3, \tag{119}$$

$$b_\nu^{(n+1)} = b_\nu^{(n)} - \delta t^{(n)} \hat{C}_\nu\left(t^{(n)}\right) \ \ \mathtt{Re}\left[\mathbf{grad}_b \mathscr{J} \frac{\partial b}{\partial b_\nu}\right]^{(n)}, \ \nu = 1, \ldots, 4. \tag{120}$$

Example: Reservoir Characterization–Parameterized Internal Profiles
In the example of history matching in reservoir engineering as discussed in section "Example 2: History Matching in Petroleum Engineering," one level set function and two interior parameter profiles need to be reconstructed from the given data, where one interior parameter profile is assumed to be smoothly varying, and the other one is assumed to overall follow a bilinear pattern. The case of smoothly varying interior parameter profiles is completely analogous to the situation

discussed in the previous section for microwave breast screening, such that it is not considered here. In the following, the situation is treated where both interior profiles follow a parameterized model with a certain set of given basis functions. In the history matching application as well as in the microwave breast screening application, the mixed cases of partly parameterized (e.g., with a constant or a bilinear profile) and partly smooth profiles are straightforward to implement as combinations of these two general approaches.

In this approach, it is assumed that the two internal profiles can be written in the parameterized form

$$b_i(\mathbf{x}) = \sum_{j=1}^{N_i} \alpha_j a_j(\mathbf{x}), \quad b_e(\mathbf{x}) = \sum_{k=1}^{N_e} \beta_k b_k(\mathbf{x}), \tag{121}$$

where a_j and b_k are the selected basis functions for each of the two domains D and $\Omega - D$, respectively. See the model discussed in section "A Modification of Color Level Set for Reservoir Characterization." In the inverse problem, the level set function ϕ and the weights α_j and β_k need to be estimated with the goal to reproduce the measured data in some sense. In order to obtain an (artificial) evolution of the unknown quantities ϕ, α_j, and β_k, the following three general evolution equations for the level set function and for the weight parameters are formulated

$$\frac{d\phi}{dt} = f(\mathbf{x}, t, \phi, \mathscr{R}), \tag{122}$$

$$\frac{d\alpha_j}{dt} = g_j(t, \phi, \mathscr{R}), \quad \frac{d\beta_k}{dt} = h_k(t, \phi, \mathscr{R}). \tag{123}$$

In the same way as before, the goal is to define the unknown terms f, g_j, and h_k such that the mismatch in the production data decreases during the evolution. For this purpose, we reformulate the cost functional now as

$$\mathscr{J}(b(\phi, \alpha_j, \beta_k)) = \frac{1}{2}\|\mathscr{R}(b(\phi, \alpha_j, \beta_k))\|^2, \tag{124}$$

where α_j denotes the weight parameters for region D and β_k denotes the weight parameters for region $\Omega - D$. Formal differentiation of this cost functional with respect to the artificial time variable t yields, in a similar way as before, the descent directions [35]

$$f_{\text{SD}}(\mathbf{x}) = -C_1 \chi_{NB}(\phi)(b_e - b_i)\mathbf{grad}_b \mathscr{J}, \tag{125}$$

$$g_{jSD}(t) = -C(\alpha_j) \int_\Omega a_j (1 - H(\phi))\mathbf{grad}_b \mathscr{J}\, d\mathbf{x}, \tag{126}$$

$$h_{kSD}(t) = -C(\beta_k) \int_\Omega b_k H(\phi)\mathbf{grad}_b \mathscr{J}\, d\mathbf{x}, \tag{127}$$

where C_1, $C(\alpha_j)$, and $C(\beta_k)$ are again positive-valued speed factors which are used for steering the speed of evolution for each of the unknowns ϕ, α_j, and β_k individually. The narrowband function $\chi_{NB}(\phi)$ is introduced for computational convenience and can be omitted if desired. For details on this narrowband formulation, see the brief discussion held in section "Shape Evolution and Shape Optimization."

7 Regularization Techniques for Shape Evolution Driven by Data Misfit

Regularization of shape evolution can be achieved by additional additive or multiplicative terms in the cost functional which control geometric terms, as discussed in Sect. 5. Alternatively, some form of regularization can be obtained by restricting the velocity fields, level set updates, or level set functions to certain classes, often without the need to introduce additional terms into the cost functional. Some of these techniques are presented in the following.

Regularization by Smoothed Level Set Updates

In the binary case (see section "The Basic Level Set Formulation for Binary Media"), a properly chosen level set function ϕ uniquely specifies a shape $D[\phi]$. This can be described by a nonlinear operator Π mapping level set functions to parameter distributions

$$\Pi(\phi)(\mathbf{x}) = \begin{cases} b_i(\mathbf{x}), & \phi(\mathbf{x}) \le 0, \\ b_e(\mathbf{x}), & \phi(\mathbf{x}) > 0. \end{cases} \tag{128}$$

We obviously have the equivalent characterization

$$\Pi(\phi)(\mathbf{x}) = b_i(\mathbf{x})\chi_D(\mathbf{x}) + b_e(\mathbf{x})(1 - \chi_D(\mathbf{x})) \tag{129}$$

where χ_D is the characteristic function of the shape D. The "level set-based residual operator" $\mathscr{T}(\phi)$ follows as

$$\mathscr{T}(\phi) = \mathscr{R}(\Pi(\phi)). \tag{130}$$

Formal differentiation by the chain rule yields

$$\mathscr{T}'(\phi) = \mathscr{R}'(\Pi(\phi))\Pi'(\phi). \tag{131}$$

The (formal) gradient direction of the least square cost functional

$$\hat{\mathscr{J}}(\phi) = \frac{1}{2}\|\mathscr{R}(b(\phi))\|_Z^2 \tag{132}$$

is then given by

$$\mathbf{grad}_{\mathscr{J}}(\phi) = \mathscr{T}'(\phi)^* \mathscr{T}(\phi), \tag{133}$$

where $\mathscr{T}'(\phi)^*$ is the L_2-adjoint of $\mathscr{T}'(\phi)$. Moreover, formally it is calculated by standard differentiation rules that

$$\Pi'(\phi) = (b_i - b_e)\delta(\phi). \tag{134}$$

Notice that, strictly speaking, the right-hand side of (131) is not an L_2-function due to the Delta distribution which is seen in (134). Nevertheless, in order to obtain practically useful expressions in a straightforward way, it is convenient to proceed with the formal considerations and, whenever necessary, to approximate the Dirac delta distribution $\delta(\phi)$ by a suitable L_2-function; see the brief discussion on this topic held in section "Shape Evolution and Shape Optimization." For example, the narrowband function $\chi_{\phi,d}(\mathbf{x})$ as defined in (151) can be used for that purpose. Then,

$$\mathscr{T}'(\phi)^* = \Pi'(\phi)^* \mathscr{R}'(\Pi(\phi))^*. \tag{135}$$

Assuming now that $\phi \in W_1(\Omega)$ with

$$W_1(\Omega) = \left\{ \phi : \phi \in L_2(\Omega),\ \nabla\phi \in L_2(\Omega),\ \frac{\partial \phi}{\partial \nu} = 0 \text{ at } \partial\Omega \right\}, \tag{136}$$

the adjoint operator $\mathscr{T}'(\phi)^*$ needs to be replaced by a new adjoint operator $\mathscr{T}'(\phi)^\circ$ which maps back from the data space into this Sobolev space $W_1(\Omega)$. Using the weighted inner product

$$\langle v, w \rangle_{W_1(\Omega)} = \alpha \langle v, w \rangle_{L_2(\Omega)} + \beta \langle \nabla v, \nabla w \rangle_{L_2(\Omega)} \tag{137}$$

with $\alpha \geq 1$ and $\beta > 0$ being carefully chosen regularization parameters, it follows

$$\mathscr{T}'(\phi)^\circ = (\alpha I - \beta\Delta)^{-1} \mathscr{T}'(\phi)^*. \tag{138}$$

The positive definite operator $(\alpha I - \beta\Delta)^{-1}$ has the effect of mapping the L_2 gradient $\mathscr{T}'(\phi)^* \mathscr{T}(\phi)$ from $L_2(\Omega)$ towards the smoother Sobolev space $W_1(\Omega)$. In fact, different choices of the weighting parameters α and β visually have the effect of "smearing out" the unregularized updates to a different degree. In particular, high-frequency oscillations or discontinuities of the updates for the level set function are removed, which yields shapes with more regular boundaries. Notice that for $\phi \in W_1(\Omega)$, the trace $\phi\big|_\Gamma$ (which is the zero level set) is only within the intermediate Sobolev space $W_{1/2}(\Gamma)$ due to the trace theorem. Therefore, the "degree of smoothness" of the reconstructed shape boundaries Γ lies somewhere in between $L_2(\Gamma)$ and $W_1(\Gamma)$.

Sometimes it is difficult or inconvenient to apply the mapping (138) to the calculated updates $\mathscr{T}'(\phi)^*\mathscr{T}(\phi)$. Then, an approximate version can be applied instead which is derived next. Denote $f_r = \mathscr{T}'(\phi)^\circ\mathscr{T}(\phi)$ and $f_d = \mathscr{T}'(\phi)^*\mathscr{T}(\phi)$. f_r can formally be interpreted as the minimizer of the cost functional

$$\hat{\mathscr{J}}(f) = \frac{\alpha - 1}{2}\|f\|_{L_2}^2 + \frac{\beta}{2}\|\nabla f\|_{L_2}^2 + \frac{1}{2}\|f - f_d\|_{L_2}^2. \tag{139}$$

In particular, the minimization process of (139) can be attempted by applying a gradient method instead of explicitly applying $(\alpha I - \beta\Delta)^{-1}$ to f_d. The gradient flow of (139) yields a modified heat (or diffusion) equation of the form

$$\begin{aligned} v_t - \beta\Delta v &= f_d - \alpha v \quad \text{for} \quad t \in [0, \tau] \\ v(0) &= f_d, \end{aligned} \tag{140}$$

with time-dependent heating term $f_d - \alpha v$, where $\hat{v} = v(\tau)$ evolves towards the minimizer f_r of (139) for $\tau \to \infty$. Practically, it turns out that a satisfactory regularization effect is achieved if instead of (140) the simplified heat equation is solved for a few time steps only:

$$\begin{aligned} v_t - \beta\Delta v &= 0 \quad \text{for} \quad t \in [0, \tau] \\ v(0) &= f_d, \end{aligned} \tag{141}$$

for τ small, using $\hat{v} = v(\tau)$ instead of f_r as update. For more details, see [44].

The above-described regularization schemes only operate on the updates (or forcing terms f in a time-dependent setting) but *not* on the level set function itself. In particular, in the case that a satisfactory solution of the shape reconstruction problem has already been achieved such that the data residuals become zero, the evolution will stop (which sometimes is desirable). In the following subsection, we will mention some alternative regularization methods where the evolution in the above-described situation would continue until an extended cost functional combining data misfit with additional geometric terms or with additional constraints on the final level set functions is minimized.

Regularization by Explicitly Penalizing Rough Level Set Functions

Instead of smoothing the updates to the level set functions, additional terms can be added to the data misfit cost functional which have the effect of penalizing certain characteristics of the level set function. For example, a Tikhonov–Philips term for the level set function can be added to (81), which will yield the minimization problem

$$\min_{\phi} \mathscr{J}(\phi) = \frac{1}{2}\|\mathscr{R}(b(\phi))\|_Z^2 + \rho(\phi), \tag{142}$$

where $\| \ \|_Z$ denotes the canonical norm in the data space Z and where ρ denotes some additional regularization term, typically involving the norm or semi-norm in the space of level set functions, for example, $\rho(\phi) = \|\nabla\phi\|^2_{L_2}$. A discussion of different choices for $\rho(\phi)$ is provided in [93]. Alternative functionals could be applied to the level set function ϕ, as, for example, Mumford–Shah, total variation, etc., which would allow for jumps in the representing level set functions.

Regularization by Smooth Velocity Fields

In the previous two subsections, regularization tools have been discussed, which are directly linked to the level set formulation of shape evolution. In section "Regularization by Smoothed Level Set Updates," smoothing operators have been applied to the updates of the level set functions (or forcing terms) which are considered as being defined on the whole domain Ω. The additional terms discussed in section "Regularization by Explicitly Penalizing Rough Level Set Functions," on the other hand, will yield additional evolutions terms which typically have to be applied directly to the describing level set functions during the shape evolution.

An alternative concept of regularizing shape evolution, which does not directly refer to an underlying level set representation of the shapes, consists in choosing function spaces for the normal velocity fields which drive the shape evolution. These velocity fields are, as such, only defined on the zero level set, i.e., on the boundaries of the given shapes (unless extension velocities are defined for a certain reason). For example, the velocity field could be taken as an element of a Sobolev space $W_1(\Gamma)$ equipped with the inner product

$$\langle v, w\rangle_{W_1(\Gamma)} = \int_\Gamma \left(\frac{\partial v}{\partial s}\frac{\partial w}{\partial s} + vw \right) ds, \tag{143}$$

where ds is the surface increment at the boundary. This leads to a postprocessing operator applied to the gradient directions which are restricted to the boundary Γ. The action of this postprocessing operator can be interpreted as mapping the given velocity field from $L_2(\Gamma)$ towards the smoother subspace $W_1(\Gamma)$, much as it was described in section "Regularization by Smoothed Level Set Updates" for the spaces $L_2(\Omega)$ and $W_1(\Omega)$. For the above given norm (143), this is modeled by a Laplace–Beltrami operator

$$-\frac{\partial^2 v}{\partial s^2} + v = f_d\big|_\Gamma. \tag{144}$$

Weighted versions of (143) and (144), with parameters α and β as in (137), can be defined as well. These operators have the effect of smoothing the velocity fields along the boundary Γ and therefore lead to regularized level set evolutions if suitable extension velocities are chosen. Alternatively, diffusion processes along the boundary can be employed for achieving a similar effect of smoothing velocity fields. For a more detailed description of various norms and the corresponding surface flows, see [17, 50, 87].

Simple Shapes and Parameterized Velocities

An even stronger way of regularizing shape evolution is to restrict the describing level set functions or the driving velocities to be members of finite-dimensional function spaces spanned by certain sets of basis functions. As basis functions, for example, polynomials, sinusoidal or exponential functions, or any other set of linearly independent functions tailored to the specific inverse problem, can be used. Closely related to this approach is also the strategy of restricting the shapes (and thereby the shape evolution) to a small set of geometric objects, as, for example, ellipsoids. See the discussion in [31] where evolution laws for a small sample of basic shapes are derived. In a related manner, [11] considers a multiscale multiregion level set technique which adaptively adjusts the support and number of basis functions for the level set representation during the shape evolution. Also related to this approach is the projection mapping strategy for shape velocities as proposed in [12].

8 Miscellaneous On-Shape Evolution

Shape Evolution and Shape Optimization

Shape evolution and shape optimization are closely related. Assume given any velocity function $F(\mathbf{x}) = \mathbf{v}(\mathbf{x}) \cdot \mathbf{n}(\mathbf{x})$ pointing into a descent direction of the cost $\mathscr{J}$, such that $\left.\frac{\partial \mathscr{J}(b)}{\partial t}\right|_{t=0} < 0$. Then, the cost will decrease in the artificial time evolution during a sufficiently small time interval $[0, \tau]$. On the practical level, the corresponding Hamilton–Jacobi-type evolution equation for the representing level set function to be solved during the time interval $[0, \tau]$ reads

$$\frac{\partial \phi}{\partial t} + F\,|\nabla \phi| = 0, \tag{145}$$

where the variables $(\mathbf{x}, t)$ have been dropped in the notation. Using, for example, a straighforward time-discretization scheme with finite differences yields

$$\frac{\phi(\tau) - \phi(0)}{\tau} + F|\nabla \phi| = 0. \tag{146}$$

Interpreting $\phi^{(n+1)} = \phi(\tau)$ and $\phi^{(n)} = \phi(0)$ yields the iteration

$$\phi^{(n+1)} = \phi^{(n)} + \tau \delta \phi^{(n)}, \quad \phi^{(0)} = \phi_0, \tag{147}$$

where τ plays the role of the step size (which might be determined by a line-search strategy) and where

$$\delta \phi^{(n)} = F|\nabla \phi^{(n)}| \tag{148}$$

for $\mathbf{x} \in \partial D$.

In the level set optimization approach, on the other hand, updates v for a level set function $\phi \to \phi + v$ are sought which reduce a given cost. Take, for example, the situation of the basic level set formulation described in section "The Basic Level Set Formulation for Binary Media." Analogously to section "Least Squares Cost Functionals and Gradient Directions," a small perturbation v then has the effect on the cost

$$\begin{aligned} \frac{d\mathscr{J}}{d\phi} v = \frac{d\mathscr{J}}{db}\frac{db}{d\phi} v = & \ \mathtt{Re}\left\langle \mathbf{grad}_b \mathscr{J}, \frac{db}{d\phi} v \right\rangle_P \\ = & \ \mathtt{Re}\left\langle \mathbf{grad}_b \mathscr{J}, (b_e(\mathbf{x}) - b_i(\mathbf{x}))\delta(\phi) v \right\rangle_P , \end{aligned} \tag{149}$$

with $\mathbf{grad}_b \mathscr{J}$ defined in (85). Apart from the term $\delta(\phi)$, this yields similar expressions for the discrete updates as (147) if choosing $F|\nabla\phi^{(n)}| = -\ \mathtt{Re}(b_e(\mathbf{x}) - b_i(\mathbf{x}))\mathbf{grad}_b \mathscr{J}$.

In fact, the term $\delta(\phi)$ is the one which causes the biggest conceptual problem when interpreting the above scheme in an optimization framework. Notice that, strictly speaking, the mapping from the level set function to the data (or to the corresponding least square cost) is not differentiable in standard (e.g., L_2) function spaces. This is indicated by the appearance of this Dirac delta distribution $\delta(\phi)$ in (149), which is not an L_2 function.

There are several ways to circumvent these difficulties, mainly aiming at replacing this troublesome Delta distribution by a better behaved approximation of it. First, in the narrowband approach, the Dirac delta is replaced by a narrow band function $\chi_{\phi,d}(\mathbf{x})$ which yields

$$F_d|\nabla\phi^{(n)}|(\mathbf{x}) = -\ \mathtt{Re}\left((b_e - b_i)\,\chi_{\phi,d}(\mathbf{x})\,\mathbf{grad}_b \mathscr{J}\right) \quad \text{for all} \quad \mathbf{x} \in \Omega . \tag{150}$$

Here, $\chi_{\phi,d}(\mathbf{x})$ is an arbitrary positive-valued approximation to $\delta(\phi)$ where the subscript d indicates the degree of approximation. For example, it can be chosen as

$$\chi_{\phi,d}(\mathbf{x}) = \begin{cases} 1\,, & \text{there exists } \mathbf{x}_0 \in \Omega \text{ with } |\mathbf{x} - \mathbf{x}_0| < d \text{ and } \phi(\mathbf{x}_0) = 0 \\ 0\,, & \text{otherwise} \end{cases} \tag{151}$$

which has the form of a "narrowband" function. Other approximations with certain additional properties (e.g., on smoothness) are possible as well. This search direction obviously also provides a descent flow for $\mathscr{J}$. In fact, the term $|\nabla\phi^{(n)}|$ can also be neglected in (150), without losing the descent property of the resulting flow, since formally it can be assumed that $|\nabla\phi^{(n)}| > 0$ (repeated recalculation of a signed distance function would even enforce $|\nabla\phi^{(n)}| = 1$).

The Dirac delta could as well be replaced by a positive constant, say 1, which yields another choice for a descent direction

$$F_{\text{top}}(\mathbf{x}) = -\ \mathtt{Re}\,(b_e - b_i)\,\mathbf{grad}_b \mathscr{J} \quad \text{for all} \quad \mathbf{x} \in \Omega . \tag{152}$$

This new direction $F_{\text{top}}(\mathbf{x})$ has the property that it applies updates driven by data sensitivities on the entire domain and thereby enables the creation of objects far away from the actual zero level set by lowering a positive level set function until its values arrive at zero. Certainly, at this moment when the level set function changes at some points far away from the zero level set from positive to negative values, a new object is created, and the descent property with respect to the cost needs to be evaluated by introducing some concept evaluating the effect of object creation on the data misfit. A formal way of doing so is briefly discussed in section "Topological Derivatives" further below. The opposite effect that inside a given shape, with some distance from the zero level set, the values of the level set function switch from negative to positive values can also occur, in which case a hole is created inside the shape. Also here, justification of this hole creation with respect to its effect on the data misfit cost is needed and can be treated as well by the tools discussed in section "Topological Derivatives."

Notice that, in the more classical level set framework, these replacements of the Dirac delta by functions with extended support can be interpreted as different ways of defining extension velocities for the numerical level set evolution scheme. Refer to [7, 19, 45, 47, 84] and the further references given there for numerical approaches which are focusing on incorporating topology changes during the shape reconstruction.

Once optimization schemes are considered for level set-based shape reconstruction, a rich set of classical optimization schemes can be adapted and applied to this novel application. For example, Newton-type optimization techniques and second-order shape derivatives can be defined and calculated. Strategies for doing so are available in the literature; see, for example, [30]. Also quasi-Newton-, Gauss–Newton-, or Levenberg–Marquardt-type schemes look promising in this framework. Some related approaches can be found, for example, in [18, 50, 82, 90, 93]. There exists a large amount of literature concerned with shape optimization problems in various applications. One important application is, for example, the structural optimal shape design problem, where the shape of a given object (a tool, bridge, telegraph pole, airplane wing, etc.) needs to be optimized subject to certain application-specific constraints [3, 87, 96]. Another example is the optimization of a bandgap structure or of maximal eigenvalues [46, 56, 75]. Some techniques from nonlinear optimization which have been successful in those applications consequently have also found their way into the treatment of shape inverse problems. For brevity, we simply refer here to the discussions presented in [2, 18, 20, 43, 50, 82, 90, 93] and the many further references therein. Alternative nonlinear algebraic reconstruction techniques are employed in [34] and fixed point techniques in [22, 23].

Some Remarks on Numerical Shape Evolution with Level Sets

Not much is said here regarding numerical schemes for solving Hamilton–Jacobi equations numerically or for solving the related optimality systems for shape inverse problems numerically. In the framework of imaging science, various schemes have

been developed and discussed extensively in the vast literature on numerical level set evolution; see, for example, the books and reviews [74, 85, 91], to mention just a few examples. These schemes include CFL conditions, re-initialization of level set functions, signed distance functions, the fast marching method, higher-order upwind schemes like ENO (essentially non-oscillating) and WENO (weighted essentially non-oscillating), artificial viscosity solutions, numerical discretizations of mean curvature terms in the level set framework, etc. All these techniques can be applied when working on the treatment of inverse problems by a level set formulation.

It is emphasized here, however, that the application of image reconstruction from indirect data comes with a number of additional problems and complications which are due to the ill-posedness of the inverse problem and to the often high complexity of the PDE (or IE) involved in the simulation of the data. Therefore, each particular image reconstruction problem from indirect data requires a careful study of numerical schemes which typically are tailor-made for the specific application. Overall, a careful choice of numerical discretization schemes and regularization parameters is indeed essential for a stable and efficient solution of the shape reconstruction problem. Moreover, also design parameters of the experimental setup (as, e.g., source and receiver locations) during the data collection have a significant impact on the shape evolution later on in the reconstruction process. Judicious choices here pay out in form of faster and more reliable reconstructions.

Speed of Convergence and Local Minima

Level set methods for shape reconstruction in inverse problems have initially been claimed to suffer from slow convergence due to inherent time-discretization constraints (the CFL condition) for the Hamilton–Jacobi equation and due to the (so far) exclusive use of first-order shape derivatives. Also, it had been observed that the shape evolution sometimes gets trapped in local minima, such that, for example, some topological components are missed by the shape evolution when starting with an inappropriate intitial guess.

However, these initial problems seem to have been resolved by now, and it appears that level set methods have in fact become quite efficient and stable when following certain straightforward guidelines and often even clearly outperform many classical pixel-based reconstruction schemes when additional prior information is available.

Firstly, the search for a good starting guess for the shape evolution can usually be done by either specific preprocessing steps (as, e.g., in [98]) or by employing more traditional search routines for only a few iteration steps. This helps avoiding "long-distance evolutions" during the succeeding shape reconstruction process.

A similar effect is achieved by the incorporation of some form of "topological derivative" in the shape evolution algorithm; see the brief discussion of this topic in the following section "Topological Derivatives." With this topological derivative technique, "seed" objects occur during the evolution just at the correct locations to be deformed in only few more iterations to their final shapes.

The topological derivative (or an appropriately designed extension velocity which has a similar effect) can also help in avoiding the shape evolution to become trapped in local minima due to barriers of low sensitivity where velocity fields become very small. Again by the effect of the creation of "seed" objects in areas of higher sensitivity, the shape evolution can jump over these barriers and quickly arrive at the final reconstruction. When an object is extended over an area of low sensitivity, then, certainly, any reconstruction scheme has difficulties with its reconstruction inside this zone, such that additional prior information might be needed for arriving at a satisfactory result inside this zone of low sensitivity (regardless which reconstruction technique is used).

In addition, also higher-order shape derivatives have been developed in the literature (see, e.g., [30]) which can be used for deriving higher-order shape-based reconstruction schemes. So far, however, their usefulness as part of a level set-based shape inversion technique has been investigated only to a very limited extent.

Finally, in an optimization framework, line-search techniques can replace the CFL condition for marching toward the sought minimum of a cost functional. This can speed up convergence significantly.

Keeping these simple strategies in mind, level set-based reconstruction techniques can in fact be much faster than more traditional schemes, in particular when the contrast value of the parameters is assumed to be known and does not need to be recovered simultaneously with the shape. For very ill-posed inverse problems, traditional techniques need a large number of iterations to converge to the right balance between correct volume and contrast value of the sought objects.

Topological Derivatives

Even though the level set formulation allows for automatic topology changes during the shape evolution, the concepts on calculating descent directions derived so far do not really apply at the moment when a topological change occurs. This is typically no problem for the case of splitting and merging of shapes, since descent directions are only calculated for discrete time steps, such that practically always never the need arises to calculate a descent direction just when such a topological change occurs. Still, from a theoretical perspective, it would be interesting to calculate expressions also for topological derivatives which capture the splitting and merging of shapes.

Another situation where topological changes occur in shape evolution is the creation and annihilation of shape components. These situations also occur automatically in the level set framework when a suitable extension velocity is chosen. However, for these two situations, explicit expressions have been derived in the literature which describe the impact of an infinitesimal topological change on the least squares data misfit cost. These are generally known as *topological derivatives*.

The technique of topological derivatives has received much attention lately as a direct way of image reconstruction. The idea in these approaches is usually to calculate a value of the topological derivative (or topological sensitivity) at each

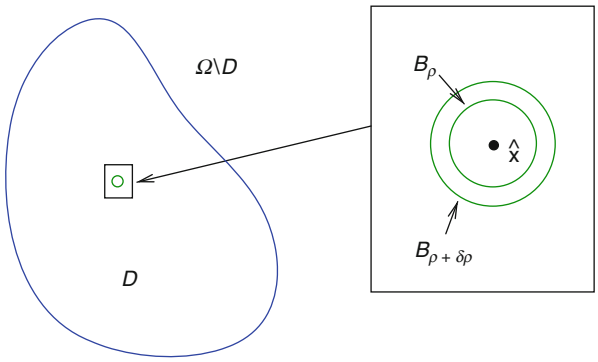


Fig. 10 Creating a hole B_ρ inside the shape D

location of the imaging domain and then adding small geometric objects at those places where this topological sensitivity is the most negative.

Certain issues arise here, as, for example, the question on how large these new objects should be, how close to each other or to the domain boundary they can be, and which contrast value should be applied for the so created small object. So far, in most cases, just one update in line with the above said is done, and thereafter the image reconstruction is either stopped or continued by a shape evolution of the so constructed set of objects. Nevertheless, the possibility of iterative topological reconstruction techniques remains an interesting challenge. Furthermore, the combination of simultaneous topological and shape evolution seems to be a very promising approach which combines the flexibility of level set evolution with the sensitivity driven creation and annihilation of shapes. This effect occurs in practice automatically if appropriate extension velocities are chosen in the regular level set shape evolution technique.

In the following, a more formal approach to topological changes is presented which has the advantage of providing a stronger mathematical justification of topological changes in the goal of data misfit reduction. The discussion will be based on the general ideas described in the references [15,22,23,39,40,69,73,83,86]. The *topological derivative* as described here aims at introducing either a small hole (let us call it B_ρ) into an existing shape D or at adding a new object (let us call it D_ρ) into the background material at some distance away from an already existing shape D (see Fig. 10). We will concentrate in the following on the first process, namely, adding a small hole into an existing shape. The complementary situation of creating a new shape component follows the same guidelines.

Denote $\tilde{D}_\rho = D \backslash B_\rho$, where the index ρ indicates the "size" of the hole B_ρ and where it is assumed that the family of new holes defined by this index is "centered" at a given point $\hat{\mathbf{x}}$. (In other words, one has $\hat{\mathbf{x}} \in B_\rho \subset B_{\rho'}$ for any $0 < \rho < \rho' < 1$.) It is assumed that all boundaries are sufficiently smooth. Consider then a cost functional $\mathscr{J}(D)$ which depends on the shape D. The topological derivative $\mathscr{D}_T$ is defined as

$$\mathscr{D}_T(\hat{\mathbf{x}}) = \lim_{\rho \downarrow 0} \frac{\mathscr{J}(\tilde{D}_\rho) - \mathscr{J}(D)}{f(\rho)}, \tag{153}$$

where $f(\rho)$ is a function which approaches zero monotonically, i.e., $f(\rho) \to 0$ for $\rho \to 0$. With this definition, the asymptotic expansion follows

$$\mathscr{J}(\tilde{D}_\rho) = \mathscr{J}(D) + f(\rho)\mathscr{D}_T(\hat{\mathbf{x}}) + o(f(\rho)). \tag{154}$$

Early applications of this technique (going back to [22, 23, 83]) were focusing on introducing ball-shaped holes into a given domain in connection to Dirichlet or Neumann problems for a Laplace equation. Here, the function $f(\rho)$ is mainly determined by geometrical factors of the created shape, and the topological derivative $\mathscr{J}(\tilde{D}_\rho)$ can be determined by solving one forward and one adjoint problem for the underlying Laplace equation. In fact, for the *Neumann problem for the Laplace equation* using ball-shaped holes, the relationship (153) takes the original form introduced in [22, 23, 83] where $f(\rho)$ is just the negative of the volume measure of the ball, i.e., $f(\rho) = -\pi\rho^2$ in 2D and $f(\rho) = -4\pi\rho^3/3$ in 3D. For more details and examples, see [22]. In general, the details of the behavior of the limit in (153), as well as of the function $f(\rho)$ if the limit exists, depend strongly on the shape of the hole, on the boundary condition at the hole interface, and on the underlying PDE.

An attempt has been made recently to find alternative definitions for the topological derivative. One such approach has been presented in [39, 40, 73]. Instead of taking the point of view that a hole is "created," the topological derivative is modeled via a limiting process where an already existing hole gradually shrinks until it disappears. For example, perturb the parameter ρ of an existing hole by a small amount $\delta\rho$. Then, the cost $\mathscr{J}(\tilde{D}_\rho)$ is perturbed to $\mathscr{J}(\tilde{D}_{\rho+\delta\rho})$, and the following limit appears:

$$\mathscr{D}_T^*(\hat{\mathbf{x}}) = \lim_{\rho \to 0} \left\{ \lim_{\delta\rho \to 0} \frac{\mathscr{J}(\tilde{D}_{\rho+\delta\rho}) - \mathscr{J}(\tilde{D}_\rho)}{f(\rho+\delta\rho) - f(\rho)} \right\}. \tag{155}$$

In [40, 73] the authors show a relationship between (153) and (155), which reads as

$$\mathscr{D}_T(\hat{\mathbf{x}}) = \mathscr{D}_T^*(\hat{\mathbf{x}}) = \lim_{\rho \to 0} \frac{1}{f'(\rho)|\mathbf{V}_n|} \mathscr{D}_{\mathbf{V}_n}(\rho), \tag{156}$$

where $\mathscr{D}_{\mathbf{V}_n}(\rho)$ is a specific form of a shape derivative related to a velocity flow $\mathbf{V}_n$ in the inward normal direction of the boundary ∂B_ρ with speed $|\mathbf{V}_n|$. For more details, refer to [40, 73]. A related link between shape derivative and topological derivative has been demonstrated also in [22]. Recently published-related work on this topic is briefly reviewed in [33].

9 Case Studies

Case Study: Microwave Breast Screening

In section "Example 1: Microwave Breast Screening," a complex breast model is presented for tackling the problem of early breast cancer detection from microwave data. Due to the high complexity of the model, also the reconstruction algorithm is likely to show some complexity. In [52] a reconstruction technique is proposed which uses five consecutive stages for the reconstruction. In the first stage, a pixel-by-pixel reconstruction is performed for the interior fatty fibroglandular region, with the skin region being (at this stage of the algorithm typically still incorrectly) estimated and fixed. Once a pixel-based reconstruction has been achieved, an initial shape for the fibroglandular region (the background being then fatty tissue) is extracted from it, which then, in the succeeding stages, is evolved jointly with the interior profiles, the skin region, and a possible tumor region until the final reconstruction is achieved. An important feature of the algorithm is that in different stages of the algorithm, different combinations of the unknowns (level set

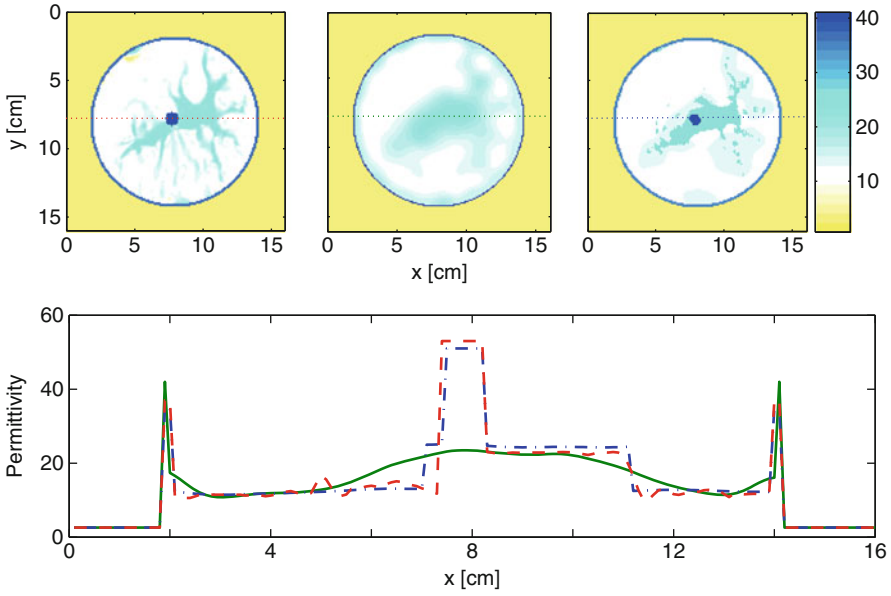


Fig. 11 First breast model of Fig. 1 with a disk-shaped tumor of diameter 8 mm situated deeply inside the breast. *Top left*: reference permittivity profile (true tumor permittivity value $\varepsilon_s^{\text{tum}} = 53$). *Top center*: the result at the end of stage I (pixel-by-pixel reconstruction). *Top right*: final reconstruction of level set-based structural inversion scheme (reconstructed permittivity value $\varepsilon_{\text{st}}^{\text{reconst}} = 50$). *Bottom*: cross section through the correct tumor for constant y coordinate (the *dashed line* represents the true permittivity profile, the *solid line* the pixel-by-pixel result, and the *dash-dotted line* the structural inversion result). For more details, see [52]

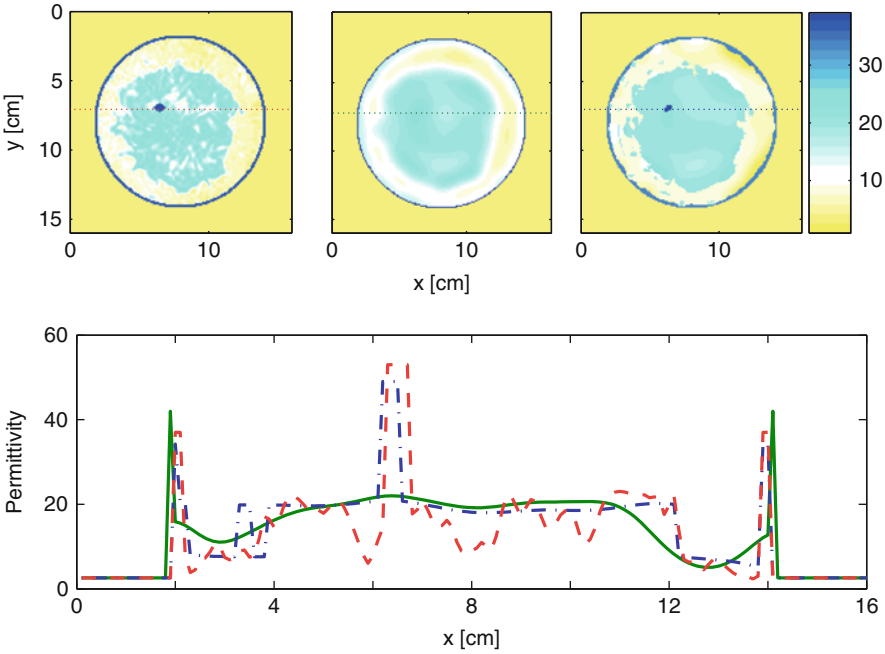


Fig. 12 Second breast model of Fig. 1 with a large fibroglandular tissue and a disk-shaped tumor of 6 mm diameter. The images are arranged as in Fig. 11. The real static permittivity of the tumor is $\varepsilon_{\mathrm{st}}^{\mathrm{tumor}} = 52$ and the reconstructed one is $\varepsilon_{\mathrm{st}}^{\mathrm{reconst}} = 49$. See also the animated movie provided in [52] which shows the shape evolution for this example

functions and interior parameter profiles) are evolved. For more details regarding the reconstruction algorithm, refer to [52].

Here, the pixel-by-pixel reconstructions of stage I of the algorithm and the final reconstructions using the complex breast model and a level set evolution are presented for the three breast models introduced in Fig. 1 and compared with each other in the cases where a small tumor is present. See Figs. 11–13. The upper left image of each figure shows the real breast, the central upper image shows the pixel-by-pixel reconstruction with our basic reconstruction scheme, and the upper right image shows the level set-based reconstruction using the complex breast model explained in section "Example 1: Microwave Breast Screening." The bottom images show cross sections through a horizontal line indicated in the upper row images and passing through the tumor locations for the three images.

The data are created on a different grid than the one used for the reconstruction. The corresponding signal-to-noise ratio is 26 dB. Forty antennas are used as sources and as receivers, which are situated equidistantly around the breast. Microwave frequencies of 1, 2, 3, 4, and 5 GHz are used for the illumination of the breast.

Even though the pixel-by-pixel reconstruction scheme is not optimized here, a general problem of pixel-based reconstruction can be identified immediately from

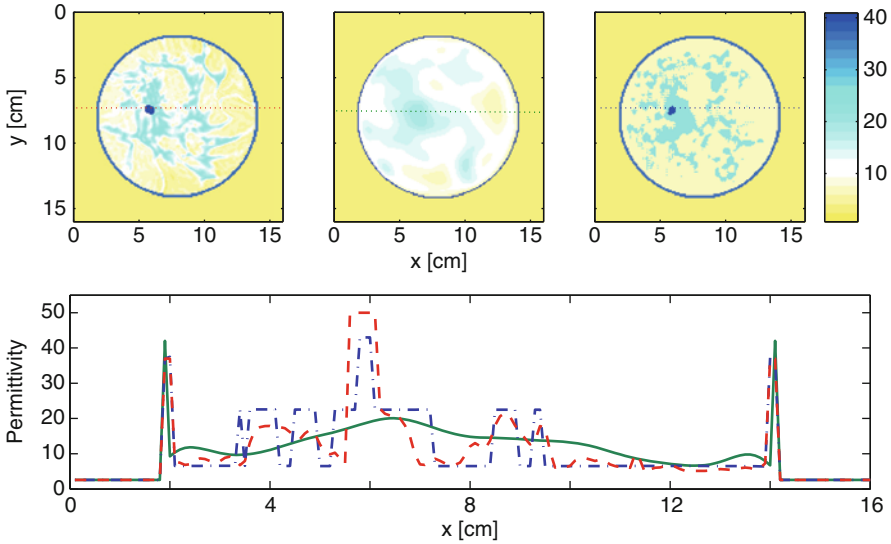


Fig. 13 Third breast model of Fig. 1 with a region of fibroglandular tissue intermixed with adipose tissue. The hidden tumor is an ellipsoid of 5×6 mm (lengths of principle axes). The images are displayed as in Fig. 11. The real static permittivity value of the tumor is $\varepsilon_{\mathrm{st}}^{\mathrm{tumor}} = 50$ and the reconstructed one is $\varepsilon_{\mathrm{st}}^{\mathrm{reconst}} = 42$. For more details, see [52]

the presented examples. The reconstructions tend to be oversmoothed, and the small tumor can hardly be identified from the pixel-based reconstruction. By no means it is possible to give any reliable estimate from these pixel-based reconstructions for the contrast of the interior tumor values to the fibroglandular or fatty tissue values of static relative permittivity. True, the level set reconstruction scheme takes advantage of the fact that it closely follows the correct model for breast tissue. On the other hand, this information is typically available (at least approximately) in breast screening applications, such that better estimates of the tumor characteristics can be expected when using such a level set-based complex breast model. This is confirmed in the three reconstructions shown in the upper right images of Figs. 11–13. For more details on this reconstruction scheme in microwave breast screening, and for an animated movie showing the image evolution, see [52].

Case Study: History Matching in Petroleum Engineering

Figure 14 shows the situation described in section "Example 2: History Matching in Petroleum Engineering" of history matching from production data. The image is composed of one zone of overall (approximately) bilinear behavior (a trend) and another zone where the permeability is smoothly varying without any clearly identifiable trend. The reconstruction follows this model and evolves simultaneously

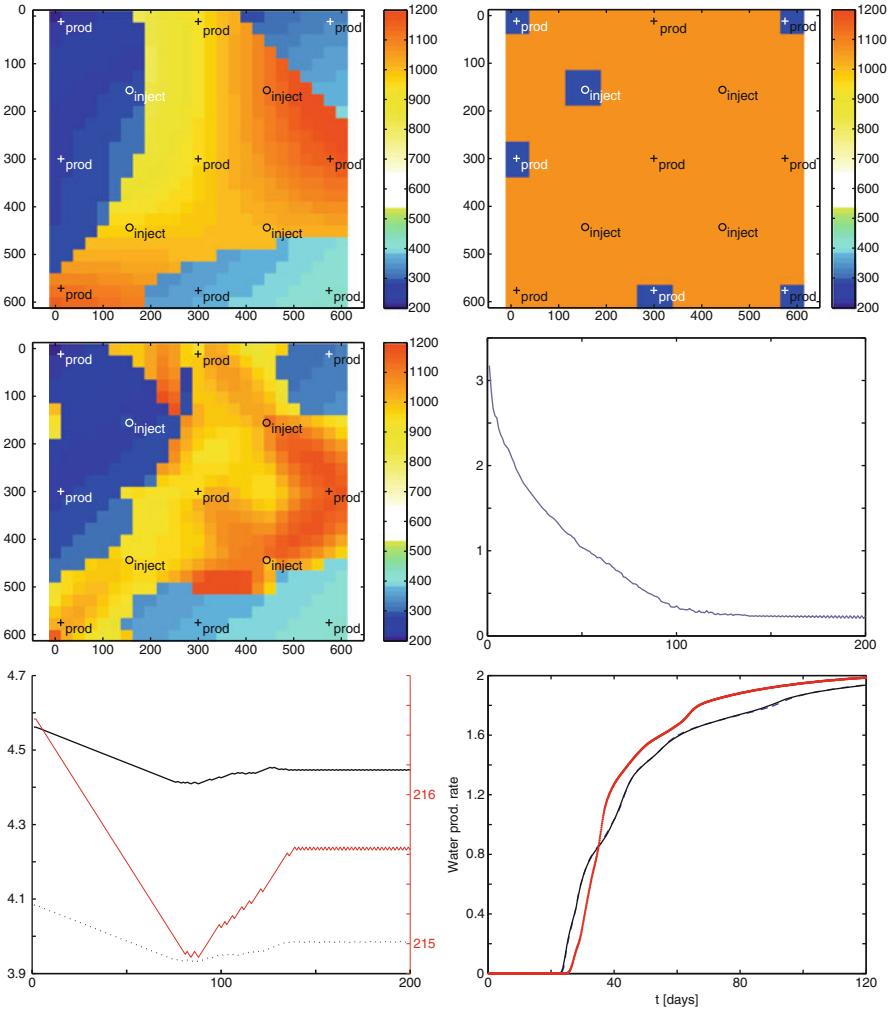


Fig. 14 Case study: history matching in reservoir engineering from production data. *Left column from top to bottom*: reference model, final reconstruction, and evolution of parameter values β_1, β_2, β_3 of the bilinear trend model; *right column from top to bottom*: initial guess, evolution of the least squares data misfit and the initial (*red solid*), final (*black dashed*), and reference (*black solid*) total water production rate in m^3/s (i.e., the true and estimated measurements). The complete evolution as an animated file and more details on the reconstruction scheme can be found in [35]

the region boundaries (i.e., the describing level set function), the three expansion parameters of the bilinear profile in the sandstone lithofacie, as well as the smoothly varying interior permeability profile in the shale region. The initial guess (upper right image of the figure) is obtained from well-log measurements. The true image is displayed in the upper left image of the figure and the final reconstruction in

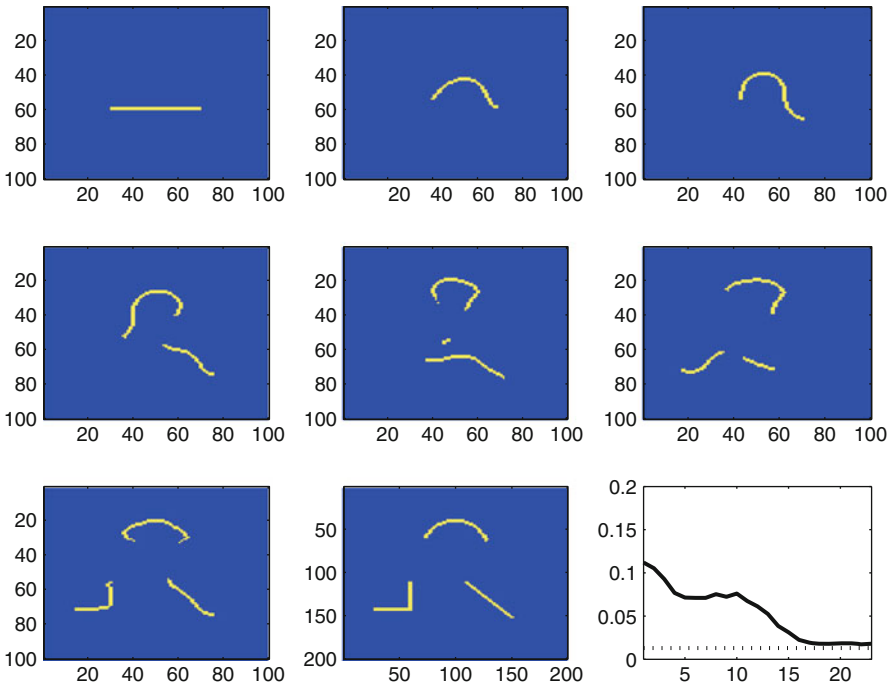


Fig. 15 Case study: crack reconstruction by an evolution of a thin shape. *Top row from left to right*: initial guess, reconstruction after 1 and 2 iterations. *Middle row*: after 5, 10, and 20 iterations. *Bottom row*: final reconstruction after 25 iterations, real crack distribution, and evolution of least squares data misfit with iteration index. The noise level is indicated in the *bottom right* image by a *horizontal dotted line*. One iteration amounts to successive application of the updates corresponding to the data of each source position in a single-step fashion

the center left image. The center right image shows the evolution of the data misfit cost during the joint evolution of all model unknowns, the lower left image shows the evolution of the three model parameters for the bilinear model in one of the regions, and the lower right image shows the initial, true, and final production rate profile over production time averaged over all boreholes (the data). A classical pixel-based reconstruction scheme typically is not able to use different models for the parameter profiles in the different regions and simultaneously reconstruct the sharp region interfaces. For more details on this reconstruction scheme for history matching in reservoir engineering, including animated movies for additional numerical examples, see [35].

Case Study: Reconstruction of Thin Shapes (Cracks)

Figure 15 shows a situation of shape evolution for the reconstruction of thin shapes (cracks) as described in section "Example 3: Crack Detection." The numerical

model presented in section "A Modification of the Classical Level Set Technique for Describing Cracks or Thin Shapes" is used here, where both level set functions describing the crack are evolved simultaneously driven by the least squares data misfit term. It is seen clearly that also in this specific model, topological changes occur automatically, when marching from a single initital (and somewhat arbitrary) crack candidate (upper left image of the figure) towards the final reconstruction showing three different crack components (bottom left image of the figure). The true situation is displayed in the bottom middle image of the figure, which shows as well three crack components which roughly are at the same location and of similar shape as the reconstructed ones. The evolution of the data misfit cost over artificial evolution time is displayed in the bottom right image of the figure. For more details on this reconstruction scheme and additional numerical experiments, see [4].

Acknowledgments OD thanks Diego Álvarez, Natalia Irishina, Miguel Moscoso and Rossmary Villegas for their collaboration on the exciting topic of level set methods in image reconstruction, and for providing figures which have been included in this chapter. He thanks the Spanish Ministerio de Educacion y Ciencia (Grants FIS2004-22546-E and FIS2007-62673), the European Union (Grant FP6-503259), the French CNRS and Univ. Paris Sud 11, and the Research Councils UK for their support of some of the work which has been presented in this chapter. DL thanks Jean Cea for having introduced him to the fascinating world of shape optimal design, Fadil Santosa for his contribution to his understanding of the linkage between shape optimal design and level set evolutions, and Jean-Paul Zolésio for his precious help on both topics, plus his many insights on topological derivatives.

Cross-References

Readers interested in the material presented in this chapter will also find interesting and relevant additional material in many other chapters of this handbook. Some additional numerical results using level set techniques can be found, for example, in the chapter on EIT. Many concepts relevant to specific implementations of level set techniques can be found, amongst others, in the following chapters.

▸ Electrical Impedance Tomography
▸ Inverse Scattering
▸ Iterative Solution Methods
▸ Large-Scale Inverse Problems in Imaging
▸ Linear Inverse Problems
▸ Optical Imaging
▸ Photoacoustic and Thermoacoustic Tomography: Image Formation Principles
▸ Regularization Methods for Ill-Posed Problems
▸ Shape Spaces
▸ Tomography
▸ Total Variation in Imaging

References

1. Abascal, J.F.P.J., Lambert, M., Lesselier, D., Dorn, O.: 3-D eddy-current imaging of metal tubes by gradient-based, controlled evolution of level sets. IEEE Trans. Magn. **44**, 4721–4729 (2009)
2. Alexandrov, O., Santosa, F.: A topology preserving level set method for shape optimization. J. Comput. Phys. **204**, 121–130 (2005)
3. Allaire, G., Jouve, F., Toader, A.-M.: Structural optimization using sensitivity analysis and a level-set method. J. Comput. Phys. **194**, 363–393 (2004)
4. Alvarez, D., Dorn, O., Irishina, N., Moscoso, M.: Crack detection using a level set strategy. J. Comput. Phys. **228**, 5710–57211 (2009)
5. Ammari, H., Calmon, P., Iakovleva, E.: Direct elastic imaging of a small inclusion. SIAM J. Imaging Sci. **1**, 169–187 (2008)
6. Ammari, H., Kang, H.: Reconstruction of Small Inhomogeneities from Boundary Measurements. Lecture Notes in Mathematics, vol. 1846. Springer, Berlin (2004)
7. Amstutz, S., Andrä, H.: A new algorithm for topology optimization using a level-set method. J. Comput. Phys. **216**, 573–588 (2005)
8. Ascher, U.M., Huang, H., van den Doel, K.: Artificial time integration. BIT Numer. Math. **47**, 3–25 (2007)
9. Bal, G., Ren, K.: Reconstruction of singular surfaces by shape sensitivity analysis and level set method. Math. Models Methods Appl. Sci. **16**, 1347–1374 (2006)
10. Ben Ameur, H., Burger, M., Hackl, B.: Level set methods for geometric inverse problems in linear elasticity. Inverse Probl. **20**, 673–696 (2004)
11. Benedetti, M., Lesselier, D., Lambert, M., Massa, A.: Multiple-shape reconstruction by means of mutliregion level sets. IEEE Trans. Geosci. Remote Sens. **48**, 2330–2342 (2010)
12. Ben Hadj Miled, M.K., Miller, E.L.: A projection-based level-set approach to enhance conductivity anomaly reconstruction in electrical resistance tomography. Inverse Probl. **23**, 2375–2400 (2007)
13. Berg, J.M., Holmstrom, K.: On parameter estimation using level sets. SIAM J. Control Optim. **37**, 1372–1393 (1999)
14. Berre, I., Lien, M., Mannseth, T.: A level set corrector to an adaptive multiscale permeability prediction. Comput. Geosci. **11**, 27–42 (2007)
15. Bonnet, M., Guzina, B.B.: Sounding of finite solid bodies by way of topological derivative. Int. J. Numer. Methods Eng. **61**, 2344–2373 (2003)
16. Burger, M.: A level set method for inverse problems. Inverse Probl. **17**, 1327–1355 (2001)
17. Burger, M.: A framework for the construction of level set methods for shape optimization and reconstruction. Interfaces Free Bound. **5**, 301–329 (2003)
18. Burger, M.: Levenberg-Marquardt level set methods for inverse obstacle problems. Inverse Probl. **20**, 259–282 (2004)
19. Burger, M., Hackl, B., Ring, W.: Incorporating topological derivatives into level set methods. J. Comput. Phys. **194**, 344–362 (2004)
20. Burger, M., Osher, S.: A survey on level set methods for inverse problems and optimal design. Eur. J. Appl. Math. **16**, 263–301 (2005)
21. Carpio, A., Rapún, M.-L.: Solving inhomogeneous inverse problems by topological derivative methods. Inverse Probl. **24**, 045014 (2008)
22. Céa, J., Garreau, S., Guillaume, P., Masmoudi, M.: The shape and topological optimizations connection. Comput. Methods Appl. Mech. Eng. **188**, 713–726 (2000)
23. Céa, J., Gioan, A., Michel, J.: Quelques résultats sur l'identification de domains. Calcolo **10**(3–4), 207–232 (1973)
24. Céa, J., Haug, E.J. (eds.): Optimization of Distributed Parameter Structures. Sijhoff & Noordhoff, Alphen aan den Rijn (1981)
25. Chan, T.F., Tai, X.-C.: Level set and total variation regularization for elliptic inverse problems with discontinuous coefficients. J. Comput. Phys. **193**, 40–66 (2003)

26. Chan, T.F., Vese, L.A.: Active contours without edges. IEEE Trans. Image Process. **10**, 266–277 (2001)
27. Chung, E.T., Chan, T.F., Tai, X.C.: Electrical impedance tomography using level set representation and total variational regularization. J. Comput. Phys. **205**, 357–372 (2005)
28. DeCezaro, A., Leitão, A., Tai, X.-C.: On multiple level-set regularization methods for inverse problems. Inverse Probl. **25**, 035004 (2009)
29. Delfour, M.C., Zolésio, J.-P.: Shape sensitivity analysis via min max differentiability. SIAM J. Control Optim. **26**, 34–86 (1988)
30. Delfour, M.C., Zolésio, J.-P.: Shapes and Geometries: Analysis, Differential Calculus and Optimization. SIAM Advances in Design and Control. SIAM, Philadelphia (2001)
31. Dorn, O., Lesselier, D.: Level set methods for inverse scattering. Inverse Probl. **22**, R67–R131 (2006). doi:10.1088/0266-5611/22/4/R01
32. Dorn, O., Lesselier, D.: Level set techniques for structural inversion in medical imaging. In: Suri, J.S., Farag, A.A. (eds.) Deformable Models, pp. 61–90. Springer, New York (2007)
33. Dorn, O., Lesselier, D.: Level set methods for inverse scattering – some recent developments. Inverse Probl. **25**, 125001 (2009). doi:10.1088/0266-5611/25/12/125001
34. Dorn, O., Miller, E., Rappaport, C.: A shape reconstruction method for electromagnetic tomography using adjoint fields and level sets. Inverse Probl. **16**, 1119–1156 (2000)
35. Dorn, O., Villegas, R.: History matching of petroleum reservoirs using a level set technique. Inverse Probl. **24**, 035015 (2008)
36. Duflot, M.: A study of the representation of cracks with level sets. Int. J. Numer. Methods Eng. **70**, 1261–1302 (2007)
37. Engl, H.W., Hanke, M., Neubauer, A.: Regularization of Inverse Problems. Mathematics and Its Applications, vol. 375. Kluwer, Dordrecht (1996)
38. Fang, W.: Multi-phase permittivity reconstruction in electrical capacitance tomography by level set methods. Inverse Probl. Sci. Eng. **15**, 213–247 (2007)
39. Feijóo, G.R.: A new method in inverse scattering based on the topological derivative. Inverse Probl. **20**, 1819–1840 (2004)
40. Feijóo, R.A., Novotny, A.A., Taroco, E., Padra, C.: The topological derivative for the Poisson problem. Math. Model Meth. Appl. Sci. **13**, 1–20 (2003)
41. Feng, H., Karl, W.C., Castanon, D.A.: A curve evolution approach to object-based tomographic reconstruction. IEEE Trans. Image Process. **12**, 44–57 (2003)
42. Ferrayé, R., Dauvignac, J.Y., Pichot, C.: An inverse scattering method based on contour deformations by means of a level set method using frequency hopping technique. IEEE Trans. Antennas Propag. **51**, 1100–1113 (2003)
43. Frühauf, F., Scherzer, O., Leitao, A.: Analysis of regularization methods for the solution of ill-posed problems involving discontinuous operators. SIAM J. Numer. Anal. **43**, 767–786 (2005)
44. González-Rodriguez, P., Kindelan, M., Moscoso, M., Dorn, O.: History matching problem in reservoir engineering using the propagation back-propagation method. Inverse Probl. **21**, 565–590 (2005)
45. Guzina, B.B., Bonnet, M.: Small-inclusion asymptotic for inverse problems in acoustics. Inverse Probl. **22**, 1761 (2006)
46. Haber, E.: A multilevel level-set method for optimizing eigenvalues in shape design problems. J. Comput. Phys. **198**, 518–534 (2004)
47. Hackl, B.: Methods for reliable topology changes for perimeter-regularized geometric inverse problems. SIAM J. Numer. Anal. **45**, 2201–2227 (2007)
48. Harabetian, E., Osher, S.: Regularization of ill-posed problems via the level set approach. SIAM J. Appl. Math. **58**, 1689–1706 (1998)
49. Hettlich, F.: Fréchet derivatives in inverse obstacle scattering. Inverse Probl. **11**, 371–382 (1995)
50. Hintermüller, M, Ring, W.: A second order shape optimization approach for image segmentation. SIAM J. Appl. Math. **64**, 442–467 (2003)
51. Hou, S., Solna, K., Zhao, H.: Imaging of location and geometry for extended targets using the response matrix. J. Comput. Phys. **199**, 317–338 (2004)

52. Irishina, N., Alvarez, D., Dorn, O., Moscoso, M.: Structural level set inversion for microwave breast screening. Inverse Probl. **26**, 035015 (2010)
53. Ito, K.: Level set methods for variational problems and application. In: Desch, W., Kappel, F., Kunisch, K. (eds.) Control and Estimation of Distributed Parameter Systems, pp. 203–217. Birkhäuser, Basel (2002)
54. Ito, K., Kunisch, K., Li, Z.: Level-set approach to an inverse interface problem. Inverse Probl. **17**, 1225–1242 (2001)
55. Jacob, M., Bresler, Y., Toronov, V., Zhang, X., Webb, A.: Level set algorithm for the reconstruction of functional activation in near-infrared spectroscopic imaging. J. Biomed. Opt. **11**, 064029 (2006)
56. Kao, C.Y., Osher, S., Yablonovitch, E.: Maximizing band gaps in two-dimentional photonic crystals by using level set methods. Appl. Phys. B **81**, 235–244 (2005)
57. Klann, E., Ramlau, R., Ring, W.: A Mumford-Shah level-set approach for the inversion and segmentation of SPECT/CT data. J. Comput. Phys. **221**, 539–557 (2008)
58. Kortschak, B., Brandstätter, B.: A FEM-BEM approach using level-sets in electrical capacitance tomography. COMPEL **24**, 591–605 (2005)
59. Leitão, A., Alves, M.M.: On level set type methods for elliptic Cauchy problems. Inverse Probl. **23**, 2207–2222 (2007)
60. Leitao, A., Scherzer, O.: On the relation between constraint regularization, level sets and shape optimization. Inverse Probl. **19**, L1–L11 (2003)
61. Lie, J., Lysaker, M., Tai, X.: A variant of the level set method and applications to image segmentation. Math. Comput. **75**, 1155–1174 (2006)
62. Lie, J., Lysaker, M., Tai, X.: A binary level set method and some applications for Mumford-Shah image segmentation. IEEE Trans. Image Process. **15**, 1171–1181 (2006)
63. Litman, A.: Reconstruction by level sets of n-ary scattering obstacles. Inverse Probl. **21**, S131–S152 (2005)
64. Litman, A., Lesselier, D., Santosa, D.: Reconstruction of a two-dimensional binary obstacle by controlled evolution of a level-set. Inverse Probl. **14**, 685–706 (1998)
65. Liu, K., Yang, X., Liu, D., et al.: Spectrally resolved three-dimensional bioluminescence tomography with a level-set strategy. J. Opt. Soc. Am. A **27**, 1413–1423 (2010)
66. Lu, Z., Robinson, B.A.: Parameter identification using the level set method. Geophys. Res. Lett. **33**, L06404 (2006)
67. Luo, Z., Tong, L.Y., Luo, J.Z., et al.: Design of piezoelectric actuators using a multiphase level set method of piecewise constants. J. Comput. Phys. **228**, 2643–2659 (2009)
68. Lysaker, M., Chan, T.F., Li, H., Tai, X.-C.: Level set method for positron emission tomography. Int. J. Biomed. Imaging **2007**, 15 (2007). doi:10.1155/2007/26950
69. Masmoudi, M., Pommier, J., Samet, B.: The topological asymptotic expansion for the Maxwell equations and some applications. Inverse Probl. **21**, 547–564 (2005)
70. Mumford, D., Shah, J.: Optimal approximation by piecewise smooth functions and associated variational problems. Commun. Pure Appl. Math. **42**, 577–685 (1989)
71. Natterer, F., Wübbeling, F.: Mathematical Methods in Image Reconstruction. Monographs on Mathematical Modeling and Computation, vol. 5. SIAM, Philadelphia (2001)
72. Nielsen, L.K., Li, H., Tai, X.C., Aanonsen, S.I., Espedal, M.: Reservoir description using a binary level set model. Comput. Vis. Sci. **13**(1), 41–58 (2008)
73. Novotny, A.A., Feijóo, R.A., Taroco, E., Padra, C.: Topological sensitivity analysis. Comput. Methods Appl. Mech. Eng. **192**, 803–829 (2003)
74. Osher, S., Fedkiw, R.: Level Set Methods and Dynamic Implicit Surfaces. Springer, New York (2003)
75. Osher, S., Santosa, F.: Level set methods for optimisation problems involving geometry and constraints I. Frequencies of a two-density inhomogeneous drum. J. Comput. Phys. **171**, 272–288 (2001)
76. Osher, S., Sethian, J.A.: Fronts propagating with curvature-dependent speed: algorithms based on Hamilton-Jacobi formulations. J. Comput. Phys. **79**, 12–49 (1988)

77. Park, W.K., Lesselier, D.: Reconstruction of thin electromagnetic inclusions by a level set method. Inverse Probl. **25**, 085010 (2009)
78. Ramananjaona, C., Lambert, M., Lesselier, D., Zolésio, J.-P.: Shape reconstruction of buried obstacles by controlled evolution of a level set: from a min-max formulation to numerical experimentation. Inverse Probl. **17**, 1087–1111 (2001)
79. Ramananjaona, C., Lambert, M., Lesselier, D., Zolésio, J.-P.: On novel developments of controlled evolution of level sets in the field of inverse shape problems. Radio Sci. **37**, 8010 (2002)
80. Ramlau, R., Ring, W.: A Mumford-Shah level-set approach for the inversion and segmentation of X-ray tomography data. J. Comput. Phys. **221**, 539–557 (2007)
81. Rocha de Faria, J., Novotny, A.A., Feijóo, R.A., Taroco, E.: First- and second-order topological sensitivity analysis for inclusions. Inverse Probl. Sci. Eng. **17**, 665–679 (2009)
82. Santosa, F.: A level set approach for inverse problems involving obstacles. ESAIM Control Optim. Calc. Var. **1**, 17–33 (1996)
83. Schumacher, A., Kobolev, V.V., Eschenauer, H.A.: Bubble method for topology and shape optimization of structures. J. Struct. Optim. **8**, 42–51 (1994)
84. Schweiger, M., Arridge, S.R., Dorn, O., Zacharopoulos, A., Kolehmainen, V.: Reconstructing absorption and diffusion shape profiles in optical tomography using a level set technique. Opt. Lett. **31**, 471–473 (2006)
85. Sethian, J.A.: Level Set Methods and Fast Marching Methods, 2nd edn. Cambridge University Press, Cambridge (1999)
86. Sokolowski, J., Zochowski, A.: On topological derivative in shape optimization. SIAM J. Control Optim. **37**, 1251–1272 (1999)
87. Sokolowski, J., Zolésio, J.-P.: Introduction to Shape Optimization: Shape Sensitivity Analysis. Springer Series in Computational Mathematics, vol. 16. Springer, Berlin (1992)
88. Soleimani, M.: Level-set method applied to magnetic induction tomography using experimental data. Res. Nondestruct. Eval. **18**(1), 1–12 (2007)
89. Soleimani, M., Dorn, O., Lionheart, W.R.B.: A narrowband level set method applied to EIT in brain for cryosurgery monitoring. IEEE Trans. Biomed. Eng. **53**, 2257–2264 (2006)
90. Soleimani, M., Lionheart, W.R.B., Dorn, O.: Level set reconstruction of conductivity and permittivity from boundary electrical measurements using experimental data. Inverse Probl. Sci. Eng. **14**, 193–210 (2005)
91. Suri, J.S., Liu, K., Singh, S., Laxminarayan, S.N., Zeng, X., Reden, L.: Shape recovery algorithms using level sets in 2D/3D medical imagery: a state-of-the-art review. IEEE Trans. Inf. Technol. Biomed. **6**, 8–28 (2002)
92. Tai, X.-C., Chan, T.F.: A survey on multiple level set methods with applications for identifying piecewise constant functions. Int. J. Numer. Anal. Model. **1**, 25–47 (2004)
93. Van den Doel, K., Ascher, U.M.: On level set regularization for highly ill-posed distributed parameter estimation problems. J. Comput. Phys. **216**, 707–723 (2006)
94. van den Doel, K., et al.: Dynamic level set regularization for large distributed parameter estimation problems. Inverse Probl. **23**, 1271–1288 (2007)
95. Vese, L.A., Chan, T.F.: A multiphase level set framework for image segmentation using the Mumford-Shah model. Int. J. Comput. Vis. **50**, 271–293 (2002)
96. Wang, M., Wang, X.: Color level sets: a multi-phase method for structural topology optimization with multiple materials. Comput. Methods Appl. Mech. Eng. **193**, 469–496 (2004)
97. Wei, P., Wang, M.Y.: Piecewise constant level set method for structural topology optimization. Int. J. Numer. Methods Eng. **78**(4), 379–402 (2009)
98. Ye, J.C., Bresler, Y., Moulin, P.: A self-referencing level-set method for image reconstruction from sparse Fourier samples. Int. J. Comput. Vis. **50**, 253–270 (2002)
99. Zhao, H.-K., Chan, T., Merriman, B., Osher, S.: A variational level set approach to multiphase motion. J. Comput. Phys. **127**, 179–195 (1996)

Part II

Inverse Problems – Case Examples

Expansion Methods

Habib Ammari and Hyeonbae Kang

Contents

1 Introduction 536
2 Electrical Impedance Tomography for Anomaly Detection 537
Physical Principles 537
Mathematical Model 538
Asymptotic Analysis of the Voltage Perturbations 539
Numerical Methods for Anomaly Detection 541
Bibliography and Open Questions 544
3 Ultrasound Imaging for Anomaly Detection 545
Physical Principles 545
Asymptotic Formulas in the Frequency Domain 545
Asymptotic Formulas in the Time Domain 547
Numerical Methods 549
Bibliography and Open Questions 554
4 Infrared Thermal Imaging 555
Physical Principles 555
Asymptotic Analysis of Temperature Perturbations 555
Numerical Methods 557
Bibliography and Open Questions 560
5 Impediography 561
Physical Principles 561
Mathematical Model 563
Substitution Algorithm 564
Bibliography and Open Questions 565
6 Magneto-Acoustic Imaging 567
Magneto-Acousto-Electrical Tomography 567

H. Ammari (✉)
Department of Mathematics and Applications, Ecole Normale Supérieure, Paris, France
e-mail: habib.ammari@ens.fr

H. Kang
Department of Mathematics, Inha University, Incheon, Korea
e-mail: hbkang@inha.ac.kr

© Springer Science+Business Media New York 2015

O. Scherzer (ed.), *Handbook of Mathematical Methods in Imaging*,
DOI 10.1007/978-1-4939-0790-8_47

Magneto-Acoustic Imaging with Magnetic Induction 570
Bibliography and Open Questions 573
7 Magnetic Resonance Elastography 573
Physical Principles 573
Mathematical Model 573
Asymptotic Analysis of Displacement Fields 575
Numerical Methods 578
Bibliography and Open Questions 579
8 Photo-Acoustic Imaging of Small Absorbers 580
Physical Principles 580
Mathematical Model 580
Reconstruction Algorithms 581
Bibliography and Open Questions 585
9 Conclusion 585
Cross-References 585
References 586

Abstract

The aim of this chapter is to review recent developments in the mathematical and numerical modeling of anomaly detection and multi-physics biomedical imaging. Expansion methods are designed for anomaly detection. They provide robust and accurate reconstruction of the location and of some geometric features of the anomalies, even with moderately noisy data. Asymptotic analysis of the measured data in terms of the size of the unknown anomalies plays a key role in characterizing all the information about the anomaly that can be stably reconstructed from the measured data. In multi-physics imaging approaches, different physical types of waves are combined into one tomographic process to alleviate deficiencies of each separate type of waves while combining their strengths. Multi-physics systems are capable of high-resolution and high-contrast imaging. Asymptotic analysis plays a key role in multi-physics modalities as well.

1 Introduction

Inverse problems in medical imaging are in their most general form ill-posed. They literally have no solution [59, 86]. If, however, in advance one has additional structural information or can supply missing information, then one may be able to determine specific features about what one wishes to image with a satisfactory resolution and accuracy. One such type of information can be that the imaging problem is to find unknown small anomalies with significantly different parameters from those of the surrounding medium. These anomalies may represent potential tumors at an early stage.

Over the last few years, an expansion technique has been developed for the imaging of such anomalies. It has proven useful in dealing with many medical imaging problems. The method relies on deriving asymptotics. Such asymptotics have been investigated in the case of the conductivity equation, the elasticity

equation, the Helmholtz equation, the Maxwell system, the wave equation, the heat equation, and the (modified) Stokes system. A remarkable feature of this method is that it allows a stable and accurate reconstruction of the location and of some geometric features of the anomalies, even with moderately noisy data. This is because the method reduces the set of admissible solutions and the number of unknowns. It can be seen as a kind of regularization in comparison with (nonlinear) iterative approaches.

Another promising technique for efficient imaging is to combine into one tomographic process different physical types of waves. Doing so, one alleviates deficiencies of each separate type of waves while combining their strengths. Again, asymptotic analysis plays a key role in the design of robust and efficient imaging techniques based on this concept of multi-waves. In the last decade or so, work on multi-physics imaging in biomedical applications has come a long way. The motivation is to achieve high-resolution and high-contrast imaging.

The objective of this chapter is threefold: (1) to provide asymptotic expansions for both internal and boundary perturbations that are due to the presence of small anomalies, (2) to apply those asymptotic formulas for the purpose of identifying the location and certain properties of the shape of the anomalies, and (3) to design efficient inversion algorithms in multi-physics modalities.

Applications of the anomaly detection and multi-physics approaches in medical imaging are described in some detail. In particular, the use of asymptotic analysis to improve a multitude of emerging imaging techniques is highlighted. These imaging modalities include electrical impedance tomography, ultrasound imaging, infrared thermography, magnetic resonance elastography, impediography, magneto-acousto-electrical tomography, magneto-acoustic tomography with magnetic induction, and photo-acoustic imaging. They can be divided into three groups: (1) those using boundary or scattering measurements such as electrical impedance tomography, ultrasound, and infrared tomographies; (2) those using internal measurements such as magnetic resonance elastography; and (3) those using boundary measurements obtained from internal perturbations of the medium such as impediography and magneto-acoustic imaging.

As it will be shown in this chapter, modalities from group (1) can only be used for anomaly detection, while those from groups (2) and (3) can provide a stable reconstruction of a distribution of physical parameters.

2 Electrical Impedance Tomography for Anomaly Detection

Physical Principles

Electrical impedance tomography uses low-frequency electric current to probe a body; the method is sensitive to changes in electrical conductivity. By injecting known amounts of current and measuring the resulting electrical potential field at points on the boundary of the body, it is possible to “invert” such data to determine the conductivity or resistivity of the region of the body probed by the

currents. This method can also be used in principle to image changes in dielectric constant at higher frequencies, which is why the method is often called "impedance" tomography rather than "conductivity" or "resistivity" tomography. However, the aspect of the method that is most fully developed to date is the imaging of conductivity/resistivity. Potential applications of electrical impedance tomography include determination of cardiac output, monitoring for pulmonary edema, and in particular screening for breast cancer.

Recently, a commercial system called TS2000 (Mirabel Medical Systems Inc., Austin, TX) has been released for adjunctive clinical uses with X-ray mammography in the diagnostic of breast cancer. The mathematical model of the TransScan can be viewed as a realistic or practical version of the general electrical impedance system. In the TransScan, a patient holds a metallic cylindrical reference electrode, through which a constant voltage of 1–2.5 V, with frequencies spanning 100 Hz–100 KHz, is applied. A scanning probe with a planar array of electrodes, kept at ground potential, is placed on the breast. The voltage difference between the hand and the probe induces a current flow through the breast, from which information about the impedance distribution in the breast can be extracted.

The use of asymptotic analysis yields a rigorous mathematical framework for the TransScan. See [30,88] for a detailed study of this electrical impedance tomography system.

Mathematical Model

Let Ω be a smooth bounded domain in $\mathbb{R}^d$, $d = 2$ or 3 and let ν_x denote the outward normal to $\partial\Omega$ at x. Suppose that the conductivity of Ω is equal to 1. Let D denote a smooth anomaly inside Ω with conductivity $0 < k \neq 1 < +\infty$. The voltage potential in the presence of the set D of conductivity anomalies is denoted by u. It is the solution to the conductivity problem

$$\begin{cases} \nabla \cdot \big(\chi(\Omega \setminus \overline{D}) + k\chi(D)\big)\nabla u = 0 \quad \text{in } \Omega, \\ \dfrac{\partial u}{\partial \nu}\Big|_{\partial\Omega} = g \quad \left(g \in L^2(\partial\Omega), \displaystyle\int_{\partial\Omega} g\, d\sigma = 0\right), \\ \displaystyle\int_{\partial\Omega} u\, d\sigma = 0, \end{cases} \tag{1}$$

where $\chi(D)$ is the characteristic function of D.

The background voltage potential U in the absence of any anomaly satisfies

$$\begin{cases} \Delta U = 0 \quad \text{in } \Omega, \\ \dfrac{\partial U}{\partial \nu}\Big|_{\partial\Omega} = g, \\ \displaystyle\int_{\partial\Omega} U\, d\sigma = 0. \end{cases} \tag{2}$$

Let $N(x,z)$ be the Neumann function for $-\Delta$ in Ω corresponding to a Dirac mass at z. That is, N is the solution to

$$\begin{cases} -\Delta_x N(x,z) = \delta_z \quad \text{in } \Omega, \\ \dfrac{\partial N}{\partial \nu_x}\Big|_{\partial\Omega} = -\dfrac{1}{|\partial\Omega|}\,, \displaystyle\int_{\partial\Omega} N(x,z)\,d\sigma(x) = 0 \quad \text{for } z \in \Omega. \end{cases} \tag{3}$$

Note that the Neumann function $N(x,z)$ is defined as a function of $x \in \overline{\Omega}$ for each fixed $z \in \Omega$.

For B a smooth bounded domain in $\mathbb{R}^d$ and $0 < k \neq 1 < +\infty$ a conductivity parameter, let $\hat{v} = \hat{v}(B,k)$ be the solution to

$$\begin{cases} \Delta\hat{v} = 0 \quad \text{in } \mathbb{R}^d \setminus \overline{B}, \\ \Delta\hat{v} = 0 \quad \text{in } B, \\ \hat{v}|_- - \hat{v}|_+ = 0 \quad \text{on } \partial B, \\ k\dfrac{\partial\hat{v}}{\partial\nu}\Big|_- - \dfrac{\partial\hat{v}}{\partial\nu}\Big|_+ = 0 \quad \text{on } \partial B, \\ \hat{v}(\xi) - \xi \to 0 \quad \text{as } |\xi| \to +\infty. \end{cases} \tag{4}$$

Here, one denotes

$$v|_\pm(\xi) := \lim_{t\to 0^+} v\left(\xi \pm t\nu_\xi\right), \quad \xi \in \partial B,$$

and

$$\frac{\partial v}{\partial \nu_\xi}\Big|_\pm(\xi) := \lim_{t\to 0^+} \left\langle \nabla v\left(\xi \pm t\nu_\xi\right), \nu_\xi \right\rangle, \quad \xi \in \partial B,$$

if the limits exist, where ν_ξ is the outward unit normal to ∂B at ξ and $\langle , \rangle$ is the scalar product in $\mathbb{R}^d$. For ease of notation, the dot will be sometimes used for the scalar product in $\mathbb{R}^d$.

Recall that $\hat{v}$ plays the role of the first-order corrector in the theory of homogenization [79].

Asymptotic Analysis of the Voltage Perturbations

In this section, an asymptotic expansion of the voltage potentials in the presence of a diametrically small anomaly with conductivity different from the background conductivity is provided.

The following theorem gives asymptotic formulas for both boundary and internal perturbations of the voltage potential that are due to the presence of a conductivity anomaly.

Theorem 1 (Voltage perturbations). *Suppose that $D = \delta B + z, \delta$ being the characteristic size of D, and let u be the solution of (1), where $0 < k \neq 1 < +\infty$. Denote by U the background solution, that is, the solution of (2).*

(i) *The following asymptotic expansion of the voltage potential on $\partial\Omega$ holds for $d = 2, 3$:*

$$u(x) \approx U(x) - \delta^d \nabla U(z) M(k, B) \nabla_z N(x, z). \tag{5}$$

Here, $N(x, z)$ is the Neumann function, that is, the solution to (3), and $M(k, B) = \left(m_{pq}\right)_{p,q=1}^{d}$ is the polarization tensor given by

$$M(k, B) := (k - 1) \int_B \nabla \hat{v}(\xi)\, d\xi, \tag{6}$$

where $\hat{v}$ is the solution to (1).

(ii) *Let w be a smooth harmonic function in Ω. The weighted boundary measurements $I_w[U]$ satisfies*

$$I_w[U] := \int_{\partial\Omega} (u - U)(x) \frac{\partial w}{\partial \nu}(x)\, d\sigma(x) \approx -\delta^d \nabla U(z) \cdot M(k, B) \nabla w(z). \tag{7}$$

(iii) *The following inner asymptotic formula holds:*

$$u(x) \approx U(z) + \delta \hat{v}\left(\frac{x - z}{\delta}\right) \cdot \nabla U(z) \quad \text{for } x \text{ near } z. \tag{8}$$

The inner asymptotic expansion (8) uniquely characterizes the shape and the conductivity of the anomaly. In fact, suppose for two Lipschitz domains B and B' and two conductivities k and k' that $\hat{v}(B, k) = \hat{v}(B', k')$ in a domain englobing B and B', then using the jump conditions satisfied by $\hat{v}(B, k)$ and $\hat{v}(B', k')$, one can easily prove that $B = B'$ and $k = k'$.

The asymptotic expansion (5) expresses the fact that the conductivity anomaly can be modeled by a dipole far away from z. It does not hold uniformly in Ω. It shows that, from an imaging point of view, the location z and the polarization tensor M of the anomaly are the only quantities that can be determined from boundary measurements of the voltage potential, assuming that the noise level is of order δ^{d+1}. It is then important to precisely characterize the polarization tensor and derive some of its properties, such as symmetry, positivity, and isoperimetric inequalities satisfied by its elements, in order to develop efficient algorithms for reconstructing conductivity anomalies of small volume.

Some important properties of the polarization tensor are listed in the next theorem.

Theorem 2 (Properties of the polarization tensor). *For $0 < k \neq 1 < +\infty$, let $M = M(k, B) = \left(m_{pq}\right)_{p,q=1}^{d}$ be the polarization tensor associated with the bounded domain B in $\mathbb{R}^d$ and the conductivity k. Then*

(i) *M is symmetric.*
(ii) *If $k > 1$, then M is positive definite, and it is negative definite if $0 < k < 1$.*
(iii) *The following isoperimetric inequalities for the polarization tensor*

$$\begin{cases} \dfrac{1}{k-1}\operatorname{trace}(M) \leq \left(d - 1 + \dfrac{1}{k}\right)|B|, \\ (k-1)\operatorname{trace}(M^{-1}) \leq \dfrac{d-1+k}{|B|}, \end{cases} \tag{9}$$

hold, where trace *denotes the trace of a matrix and $|B|$ is the volume of B.*

The polarization tensor M can be explicitly computed for disks and ellipses in the plane and balls and ellipsoids in three-dimensional space. See [25, pp. 81–89]. The formula of the polarization tensor for ellipses will be useful here. Let B be an ellipse whose semiaxes are on the x_1- and x_2-axes and of lengths a and b, respectively. Then, $M(k, B)$ takes the form

$$M(k, B) = (k-1)|B| \begin{pmatrix} \dfrac{a+b}{a+kb} & 0 \\ 0 & \dfrac{a+b}{b+ka} \end{pmatrix}. \tag{10}$$

Formula (5) shows that from boundary measurements, one can always represent and visualize an arbitrary-shaped anomaly by means of an equivalent ellipse of center z with the same polarization tensor. Further, it is impossible to extract the conductivity from the polarization tensor. The information contained in the polarization tensor is a mixture of the conductivity and the volume. A small anomaly with high conductivity and a larger anomaly with lower conductivity can have the same polarization tensor.

The bounds (9) are known as the Hashin–Shtrikman bounds. By making use of these bounds, size and thickness estimations of B can be obtained. An inclusion whose trace of the associated polarization tensor is close to the upper bound must be infinitely thin [40].

Numerical Methods for Anomaly Detection

In this section, one applies the asymptotic formula (5) for the purpose of identifying the location and certain properties of the shape of the conductivity anomalies. Two simple fundamental algorithms that take advantage of the smallness of the anoma-

lies are singled out: projection-type algorithms and multiple signal classification (MUSIC)-type algorithms. These algorithms are fast, stable, and efficient.

Detection of a Single Anomaly: A Projection-Type Algorithm

One briefly discusses a simple algorithm for detecting a single anomaly. The reader can refer to [31,73] for further details. The projection-type location search algorithm makes use of constant current sources. One wants to apply a special type of current that makes ∇U constant in D. The injection current $g = a \cdot \nu$ for a fixed unit vector $a \in \mathbb{R}^d$ yields $\nabla U = a$ in Ω.

Assume for the sake of simplicity that $d = 2$ and D is a disk. Set

$$w(y) = -(1/2\pi) \log |x - y| \quad \text{for } x \in \mathbb{R}^2 \backslash \overline{\Omega},\ y \in \Omega\ .$$

Since w is harmonic in Ω, then from (7) to (10), it follows that

$$I_w[U] \approx \frac{(k-1)|D|}{\pi(k+1)} \frac{(x-z)\cdot a}{|x-z|^2}\ , \quad x \in \mathbb{R}^2 \backslash \overline{\Omega}\ . \tag{11}$$

The first step for the reconstruction procedure is to locate the anomaly. The location search algorithm is as follows. Take two observation lines Σ_1 and Σ_2 contained in $\mathbb{R}^2 \backslash \overline{\Omega}$ given by

$$\begin{aligned} \Sigma_1 &:= \text{ a line parallel to } a, \\ \Sigma_2 &:= \text{ a line normal to } a. \end{aligned}$$

Find two points $P_i \in \Sigma_i, i = 1, 2$, so that

$$I_w[U](P_1) = 0, \quad I_w[U](P_2) = \max_{x \in \Sigma_2} |I_w[U](x)|\ .$$

From (11), one can see that the intersecting point P of the two lines

$$\Pi_1(P_1) := \{x \mid a \cdot (x - P_1) = 0\}, \tag{12}$$

$$\Pi_2(P_2) := \{x \mid (x - P_2) \text{ is parallel to } a\} \tag{13}$$

is close to the center z of the anomaly D : $|P - z| = O(\delta^2)$.

Once one locates the anomaly, the factor $|D|(k-1)/(k+1)$ can be estimated. As it has been said before, this information is a mixture of the conductivity and the volume. A small anomaly with high conductivity and larger anomaly with lower conductivity can have the same polarization tensor.

An arbitrary-shaped anomaly can be represented and visualized by means of an ellipse or an ellipsoid with the same polarization tensor. See Fig. 1.

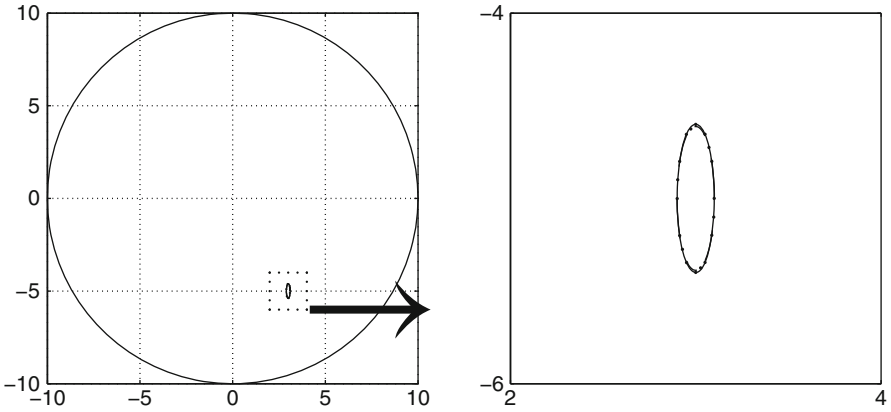


Fig. 1 Detection of the location and the polarization tensor of a small arbitrary-shaped anomaly by a projection-type algorithm. The shape of the anomaly is approximated by an ellipse with the same polarization tensor

We refer the reader to [57] for a discussion on the limits of the applicability of the projection-type location search algorithm and the derivation of a second efficient method, called the effective dipole method.

Detection of Multiple Anomalies: A MUSIC-Type Algorithm

Consider m well-separated anomalies $D_s = \delta B_s + z_s$ (these are a fixed distance apart), with conductivities k_s, $s = 1, \ldots, m$. Suppose for the sake of simplicity that all the domains B_s are disks. Let $y_l \in \mathbb{R}^2 \backslash \Omega$ for $l = 1, \ldots, n$ denote the source points. Set

$$U_{y_l} = w_{y_l} := -(1/2\pi) \log |x - y_l| \quad \text{for } x \in \Omega, \quad l = 1, \ldots, n.$$

The MUSIC-type location search algorithm for detecting multiple anomalies is as follows. For $n \in \mathbb{N}$ sufficiently large, define the response matrix $A = (A_{ll'})_{l,l'=1}^{n}$ by

$$A_{ll'} = I_{w_{y_l}}[U_{y_{l'}}] := \int_{\partial\Omega} (u - U_{y_{l'}})(x) \frac{\partial w_{y_l}}{\partial \nu}(x)\, d\sigma(x) \,.$$

Expansion (7) yields

$$A_{ll'} \approx -\sum_{s=1}^{m} \frac{2(k_s - 1)|D_s|}{k_s + 1} \nabla U_{y_{l'}}(z_s) \nabla U_{y_l}(z_s) \,.$$

Introduce

$$g(x) = \big(U_{y_1}(x), \ldots, U_{y_n}(x)\big)^* \,,$$

where v^* denotes the transpose of the vector v.

Lemma 1 (MUSIC characterization of the range of the response matrix). *There exists $n_0 > dm$ such that for any $n > n_0$, the following characterization of the location of the anomalies in terms of the range of the matrix A holds:*

$$g(x) \in \text{Range}(A) \text{ if and only if } x \in \{z_1, \ldots, z_m\} . \tag{14}$$

The MUSIC-type algorithm to determine the location of the anomalies is as follows. Let $P_{\text{noise}} = I - P$, where P is the orthogonal projection onto the range of A. Given any point $x \in \Omega$, form the vector $g(x)$. The point x coincides with the location of an anomaly if and only if $P_{\text{noise}} g(x) = 0$. Thus, one can form an image of the anomalies by plotting, at each point x, the cost function

$$W_{\text{MU}}(x) = \frac{1}{||P_{\text{noise}} g(x)||}.$$

The resulting plot will have large peaks at the locations of the anomalies.

Once one locates the anomalies, the factors $|D_s|(k_s - 1)/(k_s + 1)$ can be estimated from the significant singular values of A.

Bibliography and Open Questions

Part (i) in Theorem 1 was proven in [21, 44, 51]. The proof in [21] is based on a decomposition formula of the solution into a harmonic part and a refraction part first derived in [61]. Part (iii) is from [28]. The Hashin–Shtrikman bounds for the polarization tensor were proved in [43,77]. The projection algorithm was introduced in [31, 73]. The MUSIC algorithm was originally developed for source separation in signal theory [94]. The MUSIC-type algorithm for locating small conductivity anomalies from the response matrix was first developed in [38]. The strong relation between MUSIC and linear sampling methods was clarified in [16]. The results of this section can be generalized to the detection of anisotropic anomalies [60].

As it has been said before, it is impossible to extract separately from the detected polarization tensor information about the material property and the size of the anomaly. However, if the measurement system is very sensitive, then making use of higher-order polarization tensors yields such information. See [25] for the notion of the higher-order polarization tensors.

One of the most challenging problems in anomaly detection using electrical impedance tomography is that in practical measurements, one usually lacks exact knowledge of the boundary of the background domain. Because of this, the numerical reconstruction from the measured data is done using a model domain that represents the best guess for the true domain. However, it has been noticed that an inaccurate model of the boundary causes severe errors for the reconstructions. An elegant and original solution toward eliminating the error caused by an incorrectly modeled boundary has been proposed and implemented numerically in [69]. As nicely shown in [67], another promising approach is to use multifrequency data.

The anomaly can be detected from a weighted frequency difference of the measured boundary voltage perturbations. Moreover, this method eliminates the need for numerically simulated background measurements at the absence of the conductivity anomaly. See [58, 67].

3 Ultrasound Imaging for Anomaly Detection

Physical Principles

Ultrasound imaging is a noninvasive, easily portable, and relatively inexpensive diagnostic modality which finds extensive clinical use. The major applications of ultrasound include many aspects of obstetrics and gynecology involving the assessment of fetal health, intra-abdominal imaging of the liver and kidney, and the detection of compromised blood flow in veins and arteries.

Operating typically at frequencies between 1 and 10 MHz, ultrasound imaging produces images via the backscattering of mechanical energy from interfaces between tissues and small structures within tissue. It has high spatial resolution, particularly at high frequencies, and involves no ionizing radiation. The weaknesses of the technique include the relatively poor soft tissue contrast and the fact that gas and bone impede the passage of ultrasound waves, meaning that certain organs cannot easily be imaged. However, ultrasound imaging is a valuable technique for anomaly detection. It can be done in the time domain and the frequency domain.

Mathematical models for acoustical soundings of biological media involve the Helmholtz equation in the frequency domain and the scalar wave equation in the time domain.

Asymptotic Formulas in the Frequency Domain

Let k and ρ be positive constants. With the notation of section "Asymptotic Analysis of the Voltage Perturbations," ρ is the compressibility of the anomaly D and k is its volumetric mass density. The scalar acoustic pressure u generated by the Neumann data g in the presence of the anomaly D is the solution to the Helmholtz equation:

$$\begin{cases} \nabla \cdot \left(\chi\left(\Omega \backslash \overline{D}\right) + k\chi(D)\right) \nabla u + \omega^2 (\chi\left(\Omega \backslash \overline{D}\right) + \rho\chi(D)) u = 0 \quad \text{in } \Omega, \\ \dfrac{\partial u}{\partial \nu} = g \quad \text{on } \partial\Omega, \end{cases} \tag{15}$$

while the background solution U satisfies

$$\begin{cases} \Delta U + \omega^2 U = 0 \quad \text{in } \Omega, \\ \dfrac{\partial U}{\partial \nu} = g \quad \text{on } \partial\Omega. \end{cases} \tag{16}$$

Here, ω is the operating frequency. A relevant boundary data g is the normal derivative of a plane wave $e^{i\omega x\cdot\theta}$, with the wavelength $\lambda := 2\pi/\omega$, traveling in the direction of the unit vector θ.

Introduce the Neumann function for $-\left(\Delta+\omega^2\right)$ in Ω corresponding to a Dirac mass at z. That is, N_ω is the solution to

$$\begin{cases} -\left(\Delta_x+\omega^2\right)N_\omega(x,z)=\delta_z & \text{in } \Omega,\\ \dfrac{\partial N}{\partial \nu_x}|_{\partial\Omega}=0 & \text{on } \partial\Omega. \end{cases} \tag{17}$$

Assuming that ω^2 is not an eigenvalue for the operator $-\Delta$ in $L^2(\Omega)$ with homogeneous Neumann boundary conditions, one can prove, using the theory of relatively compact operators, existence and uniqueness of a solution to (15) at least for δ small enough [95]. Moreover, the following asymptotic formula holds.

Theorem 3 (Pressure perturbations). *Let u be the solution of (15) and let U be the background solution. Suppose that $D=\delta B+z$, with $0<(k,\rho)\neq(1,1)<+\infty$. Suppose that $\omega\delta\ll 1$.*

(i) *For any $x\in\partial\Omega$,*

$$\begin{aligned} u(x)\approx U(x)-\delta^d\big(&\nabla U(z)\cdot M(k,B)\nabla_z N_\omega(x,z)\\ &+\omega^2(\rho-1)|B|U(z)N_\omega(x,z)\big), \end{aligned} \tag{18}$$

where $M(k,B)$ is the polarization tensor associated with B and k.

(ii) *Let w be a smooth function such that $\left(\Delta+\omega^2\right)w=0$ in Ω. The weighted boundary measurements $I_w[U,\omega]$ satisfy*

$$\begin{aligned} I_w[U,\omega]&:=\int_{\partial\Omega}(u-U)(x)\frac{\partial w}{\partial\nu}(x)\,d\sigma(x)\\ &\approx -\delta^d\left(\nabla U(z)\cdot M(k,B)\nabla w(z)+\omega^2(\rho-1)|B|U(z)w(z)\right). \end{aligned} \tag{19}$$

(iii) *The following inner asymptotic formula holds:*

$$u(x)\approx U(z)+\delta\hat{v}\left(\frac{x-z}{\delta}\right)\cdot\nabla U(z)\quad \text{for } x \text{ near } z, \tag{20}$$

where $\hat{v}$ is the solution to (1).

Compared to the conductivity equation, the only extra difficulty in establishing asymptotic formulas for the Helmholtz equation (15) as the size of the acoustic anomaly goes to zero is that the equations inside and outside the anomaly are not the same.

Asymptotic Formulas in the Time Domain

Suppose that $\rho = 1$. Consider the initial boundary value problem for the (scalar) wave equation

$$\begin{cases} \partial_t^2 u - \nabla \cdot \left(\chi\left(\Omega \backslash \overline{D}\right) + k\chi(D)\right) \nabla u = 0 \quad \text{in } \Omega_T, \\ u(x,0) = u_0(x), \quad \partial_t u(x,0) = u_1(x) \quad \text{for } x \in \Omega, \\ \dfrac{\partial u}{\partial \nu} = g \quad \text{on } \partial\Omega_T, \end{cases} \tag{21}$$

where $T < +\infty$ is a final observation time, $\Omega_T = \Omega \times]0, T[$, and $\partial\Omega_T = \partial\Omega \times]0, T[$. The initial data $u_0, u_1 \in \mathcal{C}^\infty\left(\overline{\Omega}\right)$ and the Neumann boundary data $g \in \mathcal{C}^\infty(0, T; \mathcal{C}^\infty(\partial\Omega))$ are subject to compatibility conditions.

Define the background solution U to be the solution of the wave equation in the absence of any anomalies. Thus, U satisfies

$$\begin{cases} \partial_t^2 U - \Delta U = 0 \quad \text{in } \Omega_T, \\ U(x,0) = u_0(x), \quad \partial_t U(x,0) = u_1(x) \quad \text{for } x \in \Omega, \\ \dfrac{\partial U}{\partial \nu} = g \quad \text{on } \partial\Omega_T. \end{cases}$$

For $\rho > 0$, define the operator P_ρ on tempered distributions by

$$P_\rho[\psi](x,t) = \int_{|\omega| \le \rho} e^{-\sqrt{-1}\omega t} \hat{\psi}(x,\omega)\, d\omega, \tag{22}$$

where $\hat{\psi}(x,\omega)$ denotes the Fourier transform of $\psi(x,t)$ in the t-variable. Clearly, the operator P_ρ truncates the high-frequency component of ψ.

The following asymptotic expansion holds as $\delta \to 0$.

Theorem 4 (Perturbations of weighted boundary measurements). *Let $w \in C^\infty\left(\overline{\Omega}_T\right)$ satisfy $\left(\partial_t^2 - \Delta\right) w(x,t) = 0$ in Ω_T with $\partial_t w(x,T) = w(x,T) = 0$ for $x \in \Omega$. Suppose that $\rho \ll 1/\delta$. Define the weighted boundary measurements*

$$I_w[U,T] := \int_{\partial\Omega_T} P_\rho[u-U](x,t) \frac{\partial w}{\partial \nu}(x,t)\, d\sigma(x)\, dt.$$

Then, for any fixed $T > \mathrm{diam}(\Omega)$, the following asymptotic expansion for $I_w[U,T]$ holds as $\delta \to 0$:

$$I_w[U,T] \approx \delta^d \int_0^T \nabla P_\rho[U](z,t) M(k,B) \nabla w(z,t)\, dt, \tag{23}$$

where $M(k,B)$ is defined by (6).

Expansion (23) is a weighted expansion. Pointwise expansions similar to those in Theorem 1 which is for the steady-state model can also be obtained.

Let $y \in \mathbb{R}^3$ be such that $|y - z| \gg \delta$. Choose

$$U(x,t) := U_y(x,t) := \frac{\delta_{t=|x-y|}}{4\pi|x-y|} \quad \text{for } x \neq y. \tag{24}$$

It is easy to check that U_y is the outgoing Green function to the wave equation:

$$\left(\partial_t^2 - \Delta\right) U_y(x,t) = \delta_{x=y}\delta_{t=0} \quad \text{in } \mathbb{R}^3\times]0,+\infty[\,.$$

Moreover, U_y satisfies the initial conditions: $U_y(x,0) = \partial_t U_y(x,0) = 0$ for $x \neq y$. Consider now for the sake of simplicity the wave equation in the whole three-dimensional space with appropriate initial conditions:

$$\begin{cases} \partial_t^2 u - \nabla\cdot\left(\chi\left(\mathbb{R}^3\backslash\overline{D}\right) + k\chi(D)\right)\nabla u = \delta_{x=y}\delta_{t=0} \quad \text{in } \mathbb{R}^3\times]0,+\infty[\,, \\ u(x,0) = 0, \quad \partial_t u(x,0) = 0 \quad \text{for } x \in \mathbb{R}^3, x \neq y. \end{cases} \tag{25}$$

The following theorem holds.

Theorem 5 (Pointwise perturbations). *Let u be the solution to (25). Set U_y to be the background solution. Suppose that $\rho \ll 1/\delta$.*

(i) *The following outer expansion holds*

$$P_\rho[u - U_y](x,t) \approx -\delta^3 \int_{\mathbb{R}} \nabla P_\rho[U_z](x, t-\tau)\cdot M(k,B)\nabla P_\rho[U_y](z,\tau)\, d\tau\,, \tag{26}$$

for x away from z, where $M(k,B)$ is defined by (6) and U_y and U_z by (24).

(ii) *The following inner approximation holds:*

$$P_\rho[u - U_y](x,t) \approx \delta\hat{v}\left(\frac{x-z}{\delta}\right)\cdot\nabla P_\rho[U_y](x,t) \quad \text{for } x \text{ near } z, \tag{27}$$

where $\hat{v}$ is given by (4) and U_y by (24).

Formula (26) shows that the perturbation due to the anomaly is in the time domain a wave front emitted by a dipolar source located at point z.

Taking the Fourier transform of (26) in the time variable yields the expansions given in Theorem 3 for the perturbations resulting from the presence of a small anomaly for solutions to the Helmholtz equation at low frequencies (at large wavelengths compared to the size of the anomaly).

Numerical Methods

MUSIC-Type Imaging at a Single Frequency

Consider m well-separated anomalies $D_s = \delta B_s + z_s, s = 1, \ldots, m$. The compressibility and volumetric mass density of D_s are denoted by ρ_s and k_s, respectively. Suppose as before that all the domains B_s are disks. Let $(\theta_1, \ldots, \theta_n)$ be n unit vectors in $\mathbb{R}^d$. For arbitrary $\theta \in \{\theta_1, \ldots, \theta_n\}$, one assumes that one is in the possession of the boundary data u when the object Ω is illuminated with the plane wave $U(x) = e^{i\omega\theta\cdot x}$. Therefore, taking $w(x) = e^{-i\omega\theta'\cdot x}$ for $\theta' \in \{\theta_1, \ldots, \theta_n\}$ shows that one is in possession of

$$\sum_{s=1}^{m} |D_s| \left(2\frac{(k_s - 1)}{k_s + 1} \theta \cdot \theta' + (\rho_s - 1) \right) e^{i\omega(\theta - \theta')\cdot z_s},$$

for $\theta, \theta' \in \{\theta_1, \ldots, \theta_n\}$. Define the response matrix $A = (A_{ll'})_{l,l'=1}^{n} \in \mathbb{C}^{n\times n}$ by

$$A_{ll'} = \sum_{s=1}^{m} |D_s| \left(2\frac{(k_s - 1)}{k_s + 1} \theta_l \cdot \theta_{l'} + (\rho_s - 1)) \right) e^{i\omega(\theta_l - \theta_{l'})\cdot z_s}, \quad l, l' = 1, \ldots, n\ .$$

Introduce

$$g(x) = \left((1, \theta_1)^* e^{i\omega\theta_1\cdot x}, \ldots, (1, \theta_n)^* e^{i\omega\theta_n\cdot x} \right)^*.$$

Analogously to Lemma 1, one has the following characterization of the location of the anomalies in terms of the range of the matrix A.

Lemma 2 (MUSIC characterization of the range of the response matrix). *There exists $n_0 \in \mathbb{N}, n_0 > (d + 1)m$, such that for any $n \geq n_0$, the following statement holds:*

$$g^j(x) \in \text{Range}(A) \text{ if and only if } x \in \{z_1, \ldots, z_m\} \quad \text{for } j = 1, \ldots, d + 1,$$

where $g^j(x)$ is the jth column of $g(x)$.

The MUSIC algorithm can now be used as before to determine the location of the anomalies. Let $P_{\text{noise}} = I - P$, where P is the orthogonal projection onto the range of A. The imaging functional

$$W_{\text{MU}}(x) := \frac{1}{\sum_{j=1}^{d+1} ||P_{\text{noise}} g^j(x)||}$$

has large peaks only at the locations of the anomalies. See Fig. 2.

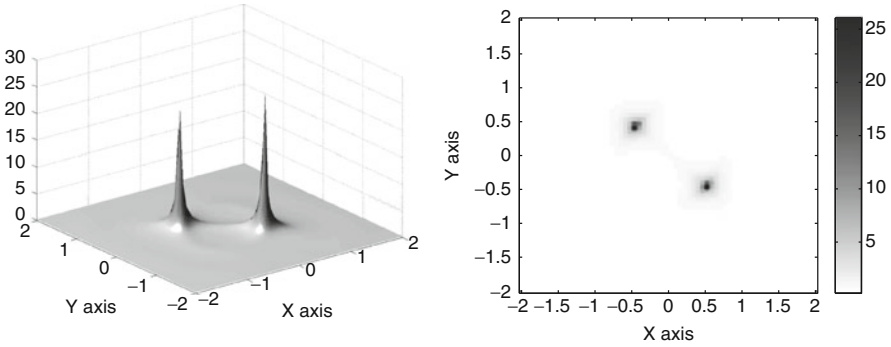


Fig. 2 MUSIC-type reconstruction from the singular value decomposition of A represented in Fig. 3

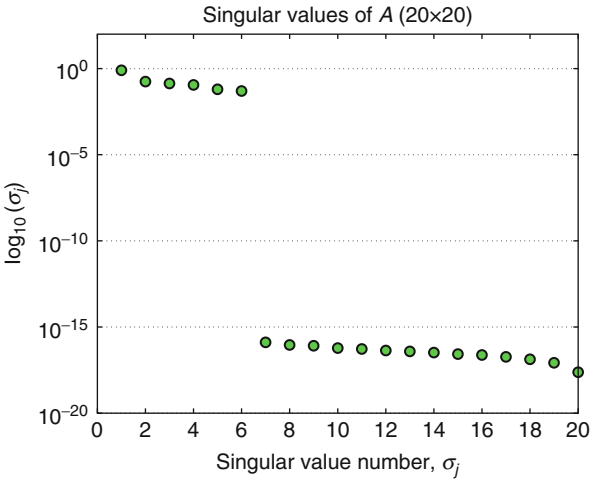


Fig. 3 Singular value decomposition of the response matrix A corresponding to two well-separated anomalies of general shape for $n = 20$, using a standard log scale. Six singular values emerge from the 14 others in the noise subspace

The significant singular vectors of A can be computed by the singular value decomposition. The number of significant singular values determines the number of anomalies. If, for example, $k_s \neq 1$ and $\rho_s \neq 1$ for all $s = 1, \ldots, m$, then there are exactly $(d + 1)m$ significant singular values of A and the rest are zero or close to zero. See Fig. 3. The significant singular values of A can be used to estimate $\frac{(k_s-1)}{k_s+1}|D_s|$ and $(\rho_s - 1)|D_s|$.

Backpropagation-Type Imaging at a Single Frequency

A backpropagation imaging functional at a single frequency ω is given by

$$W_{\mathrm{BP}}(x) := \frac{1}{n} \sum_{l=1}^{n} e^{-2i\omega\theta_l \cdot x} I_{w_l}[U_l],$$

where $U_l(x) = w_l(x) = e^{i\omega\theta_l \cdot x}$ for $\theta_l \in \{\theta_1, \ldots, \theta_n\}$. Suppose that $(\theta_1, \ldots, \theta_n)$ are equidistant points on the unit sphere S^{d-1}. For sufficiently large n, since

$$\frac{1}{n}\sum_{l=1}^{n} e^{i2\omega\theta_l \cdot x} \approx \begin{cases} j_0(2\omega|x|) & \text{for } d = 3, \\ J_0(2\omega|x|) & \text{for } d = 2, \end{cases}$$

where j_0 is the spherical Bessel function of order zero and J_0 is the Bessel function of the first kind and of order zero, it follows that

$$W_{\mathrm{BP}}(x) \approx \sum_{s=1}^{m} |D_s| \left(2\frac{(k_s - 1)}{k_s + 1} + (\rho_s - 1)\right) \times \begin{cases} j_0(2\omega|x - z_s|) & \text{for } d = 3, \\ J_0(2\omega|x - z_s|) & \text{for } d = 2. \end{cases}$$

An analogy between the backpropagation and MUSIC-type imaging can be established. Suppose that $k_s = 1$ for $s = 1, \ldots, m$. One can see that

$$W_{\mathrm{MU}}(x) \propto \frac{1}{|D_s|(\rho_s - 1) - W_{\mathrm{BP}}(x)}$$

for x near z_s [9].

Kirchhoff-Type Imaging Using a Broad Range of Frequencies

Let $y_l \in \mathbb{R}^2 \backslash \Omega$ for $l = 1, \ldots, n$ denote an array of source points. Set

$$w_{y_l}(x) = \frac{i}{4} H_0^{(1)}(\omega|x - y_l|) \quad \text{and} \quad U_{y_{l'}}(x) = \frac{i}{4} H_0^{(1)}(\omega|x - y_{l'}|),$$

where $H_0^{(1)}$ is the Hankel function of first kind and order zero. Using the asymptotic form of the Hankel function, one finds that for $\omega|x - y| \gg 1$,

$$\frac{i}{4} H_0^{(1)}(\omega|x - y|) \approx \frac{1}{2\sqrt{2\pi}} \frac{e^{i\,\pi/4}}{\sqrt{\omega|x - y|}} e^{i\omega|x-y|},$$

and

$$\frac{i}{4} \nabla H_0^{(1)}(\omega|x - y|) \approx \frac{1}{2\sqrt{2\pi}} \left(\frac{i\omega(x - y)}{|x - y|}\right) \frac{e^{i\,\pi/4}}{\sqrt{\omega|x - y|}} e^{i\omega|x-y|}.$$

Assume a high-frequency regime with $\omega L \gg 1$ for L the distance from the array center point to the locations $z_s, s = 1, \ldots, m$. It follows that

$$I_{w_l}[U_{l'}, \omega] \propto$$
$$\sum_{s=1}^{m} |D_s| \left(-2\frac{k_s - 1}{k_s + 1} \frac{(z_s - y_l) \cdot (z_s - y_{l'})}{|z_s - y_l||z_s - y_{l'}|} + (\rho_s - 1)\right) e^{i\omega(|z_s - y_l| + |z_s - y_{l'}|)}.$$

Introduce the response matrix $A(\omega) = (A_{ll'}(\omega))$ by

$$A_{ll'}(\omega) := I_{w_l}[U_{l'}, \omega]$$

and the illumination vector

$$g(x,\omega) := \left(\left(1, \frac{x-y_1}{|x-y_1|}\right)^* e^{i\omega|x-y_1|}, \ldots, \left(1, \frac{x-y_n}{|x-y_n|}\right)^* e^{i\omega|x-y_n|} \right)^*.$$

In the case of measurements at multiple frequencies (ω_j), we construct the weighted Kirchhoff imaging functional as

$$W_{\mathrm{KI}}(x) = \frac{1}{J} \sum_{\omega_j, j=1,\ldots,J} \sum_l \left(g(x,\omega_j), u_l(\omega_j)\right) \left(g(x,\omega_j), \overline{v_l}(\omega_j)\right),$$

where $(a,b) = \overline{a} \cdot b$, J is the number of frequencies, and u_l and v_l are, respectively, the left and right singular vectors of A. As for W_{MU}, W_{KI} is written in terms of the singular-value decompositions of the response matrices $A(\omega_j)$.

Time-Reversal Imaging

Unlike the three previous imaging methods, the one in this section is in time domain. It is based on time reversal.

The main idea of time reversal is to take advantage of the reversibility of the wave equation in a non-dissipative unknown medium in order to back propagate signals to the sources that emitted them. In the context of anomaly detection, one measures the perturbation of the wave on a closed surface surrounding the anomaly and retransmits it through the background medium in a time-reversed chronology. Then the perturbation will travel back to the location of the anomaly. One can show that the time-reversal perturbation focuses on the location z of the anomaly with a focal spot size limited to one-half the wavelength which is in agreement with the Rayleigh resolution limit.

In mathematical terms, suppose that one is able to measure the perturbation $u - U_y$ and its normal derivative at any point x on a sphere S englobing the anomaly D and for a large time t_0. The time-reversal operation is described by the transform $t \mapsto t_0 - t$. Both the perturbation and its normal derivative on S are time reversed and emitted from S. Then a time-reversed perturbation propagates inside the volume surrounded by S.

To detect the anomaly from measurements of the wavefield $u - U_y$ away from the anomaly, one can use a time-reversal technique. Taking into account the definition of the outgoing fundamental solution (24) to the wave equation, spatial reciprocity, and time-reversal invariance of the wave equation, one defines the time-reversal imaging functional W_{TR} by

$$W_{\mathrm{TR}}(x,t) = \int_{\mathbb{R}}\int_{S}\Bigg[U_x\left(x',t-s\right)\frac{\partial P_\rho\left[u-U_y\right]}{\partial \nu}\left(x',t_0-s\right) - \frac{\partial U_x}{\partial \nu}\left(x',t-s\right)P_\rho\left[u-U_y\right]\left(x',t_0-s\right)\Bigg]\,d\sigma\left(x'\right)\,ds, \tag{28}$$

where

$$U_x\left(x',t-\tau\right) = \frac{\delta\left(t-\tau-|x-x'|\right)}{4\pi|x-x'|}.$$

The imaging functional W_{TR} corresponds to propagating inside the volume surrounded by S, the time-reversed perturbation $P_\rho[u-U_y]$, and its normal derivative on S. Theorem 5 shows that

$$P_\rho\left[u-U_y\right](x,t) \approx -\delta^3\int_{\mathbb{R}}\nabla P_\rho[U_z](x,t-\tau)\cdot m(z,\tau)\,d\tau,$$

where

$$m(z,\tau) = M(k,B)\nabla P_\rho[U_y](z,\tau)\ . \tag{29}$$

Therefore, since

$$\begin{aligned}\int_{\mathbb{R}}\int_{S}\Bigg[&U_x(x',t-s)\frac{\partial P_\rho[U_z]}{\partial \nu}(x',t_0-s-\tau) \\ &-\frac{\partial U_x}{\partial \nu}(x',t-s)P_\rho[U_z](x',t_0-s-\tau)\Bigg]\,d\sigma(x')\,ds \\ &= P_\rho[U_z](x,t_0-\tau-t) - P_\rho[U_z](x,t-t_0+\tau),\end{aligned} \tag{30}$$

one obtains the approximation

$$W_{\mathrm{TR}}(x,t) \approx -\delta^3\int_{\mathbb{R}} m(z,\tau)\cdot\nabla_z\left[P_\rho[U_z](x,t_0-\tau-t) - P_\rho[U_z](x,t-t_0+\tau)\right]d\tau,$$

which can be interpreted as the superposition of incoming and outgoing waves, centered on the location z of the anomaly. Since

$$P_\rho[U_y](x,\tau) = \frac{\sin\rho(\tau-|x-y|)}{2\pi(\tau-|x-y|)|x-y|},$$

$m(z,\tau)$ is concentrated at the travel time $\tau = T = |z-y|$. It then follows that

$$W_{\mathrm{TR}}(x,t) \approx -\delta^3 m(z,T)\cdot\nabla_z\left[P_\rho[U_z](x,t_0-T-t) - P_\rho[U_z](x,t-t_0+T)\right]. \tag{31}$$

The imaging functional W_{TR} is clearly the sum of incoming and outgoing polarized spherical waves.

Approximation (31) has an important physical interpretation. By changing the origin of time, T can be set to 0 without loss of generality. Then by taking a Fourier transform of (31) over the time variable t, one obtains that

$$\hat{W}_{\mathrm{TR}}(x,\omega) \propto \delta^3 m(z,T)\cdot\nabla j_0(\omega|x-z|),$$

where ω is the wave number. This shows that the time-reversal perturbation W_{TR} focuses on the location z of the anomaly with a focal spot size limited to one-half the wavelength.

An identity parallel to (30) can be derived in the frequency domain. In fact, one has

$$\begin{aligned}\int_S \left[\hat{U}_x\left(x'\right)\frac{\partial\overline{\widehat{U}_z}}{\partial\nu}\left(x'\right)-\frac{\partial\hat{U}_x}{\partial\nu}\left(x'\right)\overline{\widehat{U}_z}\left(x'\right)\right]d\sigma\left(x'\right) &= 2i\,\Im m\,\widehat{U}_z(x)\\ &\propto j_0(\omega|x-z|),\end{aligned} \tag{32}$$

which shows that in the frequency domain, W_{TR} coincides with W_{BP}.

Bibliography and Open Questions

The initial boundary-value problems for the wave equation in the presence of anomalies of small volume have been considered in [5, 24]. Theorem 5 is from [12]. In [12], a time-reversal approach was also designed for locating the anomaly from the outer expansion (26). The physics literature on time reversal is quite rich. One refers, for instance, to [48] and the references therein. See [93] for clinical applications of time reversal. Many interesting mathematical works have dealt with different aspects of time-reversal phenomena: see, for instance, [33] for time reversal in the time domain, [45–47, 82] for time reversal in the frequency domain, and [37, 50] for time reversal in random media.

The MUSIC-type algorithm for locating small acoustic or electromagnetic anomalies from the multi-static response matrix at a fixed frequency was developed in [18]. See also [18–20], where a variety of numerical results was presented to highlight its potential and its limitation. It is worth mentioning that the MUSIC-type algorithm is related to time reversal [82, 87].

MUSIC and Kirchhoff imaging functionals can be extended to the time domain in order to detect the anomaly and its polarization tensor from (dynamical) boundary measurements [24].

The inner expansion in Theorem 5 can be used to design an efficient optimization algorithm for reconstructing the shape and the physical parameter of an anomaly from the near-field perturbations of the wavefield, which can be used in radiation force imaging.

In radiation force imaging, one uses the acoustic radiation force of an ultrasonic focused beam to remotely generate mechanical vibrations in organs. A spatiotemporal sequence of the propagation of the induced transient wave can be acquired, leading to a quantitative estimation of the physical parameters of the anomaly. See, for instance, [35, 36].

The proposed location search algorithms using transient wave or broad-range multifrequency boundary measurements can be extended to the case with limited-view measurements. Using the geometrical control method [34], one can still exploit those algorithms and perform imaging with essentially the same resolution using partial data as using complete data, provided that the geometric optics condition holds.

An identity similar to (32) can be derived in an inhomogeneous medium, which shows that the sharper the behavior of the imaginary part of the Green function around the location of the anomaly is, the higher is the resolution. It would be quite challenging to explicitly see how this behavior depends on the heterogeneity of the surrounding medium. This would yield super-resolved ultrasound imaging systems.

4 Infrared Thermal Imaging

Physical Principles

Infrared thermal imaging is becoming a common screening modality in the area of breast cancer. By carefully examining the aspects of temperature and blood vessels of the breasts in thermal images, signs of possible cancer or precancerous cell growth may be detected up to 10 years prior to being discovered using any other procedure. This provides the earliest detection of cancer possible.

Because of thermal imaging's extreme sensitivity, these temperature variations and vascular changes may be among the earliest signs of breast cancer and/or a precancerous state of the breast. An abnormal infrared image of the breast is an important marker of high risk for developing breast cancer. See [3, 83].

Asymptotic Analysis of Temperature Perturbations

Suppose that the background Ω is homogeneous with thermal conductivity 1 and that the anomaly $D = \delta B + z$ has thermal conductivity $0 < k \neq 1 < +\infty$. In this section, one considers the following transmission problem for the heat equation:

$$\begin{cases} \partial_t u - \nabla \cdot \left(\chi\left(\Omega \backslash \overline{D}\right) + k\chi(D)\right) \nabla u = 0 \quad \text{in } \Omega_T, \\ u(x,0) = u_0(x) \quad \text{for } x \in \Omega, \\ \dfrac{\partial u}{\partial \nu} = g \quad \text{on } \partial\Omega_T, \end{cases} \tag{33}$$

where the Neumann boundary data g and the initial data u_0 are subject to a compatibility condition. Let U be the background solution defined as the solution of

$$\begin{cases} \partial_t U - \Delta U = 0 & \text{in } \Omega_T, \\ U(x,0) = u_0(x) & \text{for } x \in \Omega, \\ \dfrac{\partial U}{\partial \nu} = g & \text{on } \partial\Omega_T. \end{cases}$$

The following asymptotic expansion holds as $\delta \to 0$.

Theorem 6 (Perturbations of weighted boundary measurements). *Let $w \in C^\infty(\overline{\Omega}_T)$ be a solution to the adjoint problem, namely, satisfy $(\partial_t + \Delta)w(x,t) = 0$ in Ω_T with $w(x,T) = 0$ for $x \in \Omega$. Define the weighted boundary measurements*

$$I_w[U,T] := \int_{\partial\Omega_T} (u-U)(x,t) \frac{\partial w}{\partial \nu}(x,t)\, d\sigma(x)\, dt.$$

Then, for any fixed $T > 0$, the following asymptotic expansion for $I_w[U,T]$ holds as $\delta \to 0$:

$$I_w[U,T] \approx -\delta^d \int_0^T \nabla U(z,t) \cdot M(k,B) \nabla w(z,t)\, dt, \tag{34}$$

where $M(k,B)$ is defined by (6).

Note that (34) holds for any fixed positive final time T, while (23) holds only for $T > \operatorname{diam}(\Omega)$. This difference comes from the finite speed propagation property for the wave equation compared to the infinite one for the heat equation.

Consider now the background solution to be the Green function of the heat equation at y:

$$U(x,t) := U_y(x,t) := \begin{cases} \dfrac{e^{-\frac{|x-y|^2}{4t}}}{(4\pi t)^{d/2}} & \text{for } t > 0, \\ 0 \quad \text{for } t < 0\,. \end{cases} \tag{35}$$

Let u be the solution to the following heat equation with an appropriate initial condition:

$$\begin{cases} \partial_t u - \nabla \cdot \left(\chi\left(\mathbb{R}^d \setminus \overline{D}\right) + k\chi(D)\right) \nabla u = 0 & \text{in } \mathbb{R}^d \times]0, +\infty[\,, \\ u(x,0) = U_y(x,0) & \text{for } x \in \mathbb{R}^d. \end{cases} \tag{36}$$

Proceeding as in the derivation of (26), one can prove that $\delta u(x,t) := u - U$ is approximated by

$$-(k-1)\int_0^t \frac{1}{(4\pi(t-\tau))^{d/2}} \int_{\partial D} e^{-\frac{|x-x'|^2}{4(t-\tau)}} \frac{\partial \hat{v}}{\partial \nu}\bigg|_- \left(\frac{x'-z}{\delta}\right) \cdot \nabla U_y\left(x',\tau\right)\, d\sigma\left(x'\right)\, d\tau, \tag{37}$$

for x near z. Therefore, analogously to Theorem 5, the following pointwise expansion follows from the approximation (37).

Theorem 7 (Pointwise perturbations). *Let $y \in \mathbb{R}^d$ be such that $|y-z| \gg \delta$. Let u be the solution to (36). The following expansion holds*

$$(u-U)(x,t) \approx -\delta^d \int_0^t \nabla U_z(x,t-\tau) M(k,B) \nabla U_y(z,\tau)\, d\tau \quad \text{for } |x-z| \gg O(\delta), \tag{38}$$

where $M(k,B)$ is defined by (6) and U_y and U_z by (35).

When comparing (38) and (26), one should point out that for the heat equation the perturbation due to the anomaly is accumulated over time.

An asymptotic formalism for the realistic half-space model for thermal imaging, well suited for the design of anomaly reconstruction algorithms, has been developed in [29].

Numerical Methods

In this section, the formula (34) is applied (with an appropriate choice of test functions w and background solutions U) for the purpose of identifying the location of the anomaly D. The first algorithm makes use of constant heat flux, and, not surprisingly, it is limited in its ability to effectively locate multiple anomalies.

Using many heat sources, one then describes an efficient method to locate multiple anomalies and illustrate its feasibility. For the sake of simplicity, only the two-dimensional case will be considered.

Detection of a Single Anomaly

For $y \in \mathbb{R}^2 \backslash \overline{\Omega}$, let

$$w(x,t) = w_y(x,t) := \frac{1}{4\pi(T-t)} e^{-\frac{|x-y|^2}{4(T-t)}}. \tag{39}$$

The function w satisfies $(\partial_t + \Delta)w = 0$ in Ω_T and the final condition $w|_{t=T} = 0$ in Ω.

Suppose that there is only one anomaly $D = z + \delta B$ with thermal conductivity k. For simplicity, assume that B is a disk. Choose the background solution $U(x,t)$ to be a harmonic (time-independent) function in Ω_T. One computes

$$\nabla w_y(z,t) = \frac{y-z}{8\pi(T-t)^2} e^{-\frac{|z-y|^2}{4(T-t)}},$$
$$M(k,B)\nabla w_y(z,t) = \frac{(k-1)|B|}{k+1} \frac{y-z}{4\pi(T-t)^2} e^{-\frac{|z-y|^2}{4(T-t)}},$$

and

$$\int_0^T M(k,B)\nabla w_y(z,t)\,dt = \frac{(k-1)|B|}{k+1} \frac{y-z}{4\pi} \int_0^T \frac{e^{-\frac{|z-y|^2}{4(T-t)}}}{(T-t)^2}\,dt.$$

But

$$\frac{d}{dt} e^{-\frac{|z-y|^2}{4(T-t)}} = \frac{-|z-y|^2}{4} \frac{e^{-\frac{|z-y|^2}{4(T-t)}}}{(T-t)^2}$$

and therefore

$$\int_0^T M(k,B)\nabla w_y(z,t)\,dt = \frac{(k-1)|B|}{k+1} \frac{y-z}{\pi|z-y|^2} e^{-\frac{|z-y|^2}{4(T-t)}}.$$

Then the asymptotic expansion (34) yields

$$I_w[U,T](y) \approx \delta^2 \frac{k-1}{k+1} |B| \frac{\nabla U(z)\cdot(y-z)}{\pi|y-z|^2} e^{-\frac{|y-z|^2}{4T}}. \tag{40}$$

Now, one is in a position to present the projection-type location search algorithm for detecting a single anomaly. Prescribe the initial condition $u_0(x) = a \cdot x$ for some fixed unit constant vector a and choose $g = a \cdot \nu$ as an applied time-independent heat flux on $\partial\Omega_T$, where a is taken to be a coordinate unit vector. Take two observation lines Σ_1 and Σ_2 contained in $\mathbb{R}^2 \setminus \overline{\Omega}$ such that

$$\Sigma_1 := \text{a line parallel to } a, \qquad \Sigma_2 := \text{a line normal to } a\ .$$

Next, find two points $P_i \in \Sigma_i (i = 1, 2)$ so that $I_w(T)(P_1) = 0$ and

$$I_w(T)(P_2) = \begin{cases} \min\limits_{x\in\Sigma_2} I_w(T)(x) & \text{if } k-1 < 0, \\ \max\limits_{x\in\Sigma_2} I_w(T)(x) & \text{if } k-1 > 0. \end{cases}$$

Finally, draw the corresponding lines $\Pi_1(P_1)$ and $\Pi_2(P_2)$ given by (12). Then the intersecting point P of $\Pi_1(P_1) \cap \Pi_2(P_2)$ is close to the anomaly $D : |P - z| = O(\delta\,|\log\delta|)$ for δ small enough.

Detection of Multiple Anomalies: A MUSIC-Type Algorithm

Consider m well-separated anomalies $D_s = \delta B_s + z_s, s = 1, \ldots, m$, whose heat conductivity is k_s. Choose

$$U(x,t) = U_{y'}(x,t) := \frac{1}{4\pi t} e^{-\frac{|x-y'|^2}{4t}} \quad \text{for } y' \in \mathbb{R}^2 \backslash \overline{\Omega}$$

or, equivalently, g to be the heat flux corresponding to a heat source placed at the point source y' and the initial condition $u_0(x) = 0$ in Ω, to obtain that

$$I_w[U,T] \approx -\delta^2 \sum_{s=1}^{m} \frac{(1-k_s)}{64\pi^2} (y' - z_s) M^{(s)} (y - z_s) \times \int_0^T \frac{1}{t^2(T-t)^2} \exp\left(-\frac{|y-z_s|^2}{4(T-t)} - \frac{|y'-z_s|^2}{4t} \right) dt,$$

where w is given by (39) and $M^{(s)}$ is the polarization tensor of D_s.

Suppose for the sake of simplicity that all the domains B_s are disks. Then it follows from (10) that $M^{(s)} = m^{(s)} I_2$, where $m^{(s)} = 2(k_s - 1)|B_s|/(k_s + 1)$ and I_2 is the 2×2 identity matrix. Let $y_l \in \mathbb{R}^2 \backslash \overline{\Omega}$ for $l \in \mathbb{N}$ be the source points. One assumes that the countable set $\{y_l\}_{l \in \mathbb{N}}$ has the property that any analytic function which vanishes in $\{y_l\}_{l \in \mathbb{N}}$ vanishes identically.

The MUSIC-type location search algorithm for detecting multiple anomalies is as follows. For $n \in \mathbb{N}$ sufficiently large, define the matrix $A = [A_{ll'}]_{l,l'=1}^{n}$ by

$$A_{ll'} := -\delta^2 \sum_{s=1}^{m} \frac{(1-k_s)}{64\pi^2} m^{(s)} (y_{l'} - z_s) \cdot (y_l - z_s) \times \int_0^T \frac{1}{t^2(T-t)^2} \exp\left(-\frac{|y_l - z_s|^2}{4(T-t)} - \frac{|y_{l'} - z_s|^2}{4t} \right) dt \,.$$

For $z \in \Omega$, one decomposes the symmetric real matrix C defined by

$$C := \left[\int_0^T \frac{1}{t^2(T-t)^2} \exp\left(-\frac{|y_l - z|^2}{4(T-t)} - \frac{|y_{l'} - z|^2}{4t} \right) dt \right]_{l,l'=1,\ldots,n}$$

as follows:

$$C = \sum_{l=1}^{p} v_l(z) v_l(z)^* \tag{41}$$

for some $p \leq n$, where $v_l \in \mathbb{R}^n$ and v_l^* denotes the transpose of v_l. Define the vector $g_z^{(l)} \in \mathbb{R}^{n \times 2}$ for $z \in \Omega$ by

$$g_z^{(l)} = ((y_1 - z)v_{l1}(z), \ldots, (y_n - z)v_{ln}(z))^*, \quad l = 1, \ldots, p\,. \tag{42}$$

Here, $v_{l1}, \ldots, v_{ln}$ are the components of the vector $v_l, l = 1, \ldots, p$. Let $y_l = (y_{lx}, y_{ly})$ for $l = 1, \ldots, n, z = (z_x, z_y)$, and $z_s = (z_{sx}, z_{sy})$. One also introduces

$$g_{zx}^{(l)} = ((y_{1x} - z_x)v_{l1}(z), \ldots, (y_{nx} - z_x)v_{ln}(z))^*$$

and

$$g_{zy}^{(l)} = \big((y_{1y} - z_y)v_{l1}(z), \ldots, (y_{ny} - z_y)v_{ln}(z)\big)^*.$$

Lemma 3 (MUSIC characterization of the range of the response matrix). *The following characterization of the location of the anomalies in terms of the range of the matrix A holds:*

$$g_{zx}^{(l)} \text{ and } g_{zy}^{(l)} \in \text{Range}(A) \ \ \forall\, l \in \{1, \ldots, p\} \quad \text{if and only if} \quad z \in \{z_1, \ldots, z_m\}\,. \tag{43}$$

Note that the smallest number n which is sufficient to efficiently recover the anomalies depends on the (unknown) number m. This is the main reason for taking n sufficiently large. As for the electrical impedance imaging, the MUSIC-type algorithm for the thermal imaging is as follows. Compute P_{noise}, the projection onto the noise space, by the singular value decomposition of the matrix A. Compute the vectors v_l by (41). Form an image of the locations, $z_1, \ldots, z_m$, by plotting, at each point z, the quantity $\left\| g_z^{(l)} \cdot a \right\| / \left\| P_{\text{noise}} \left(g_z^{(l)} \cdot a \right) \right\|$ for $l = 1, \ldots, p$, where $g_z^{(l)}$ is given by (42) and a is a unit constant vector. The resulting plot will have large peaks at the locations of $z_s, s = 1, \ldots, m$.

The next two figures (Figs. 4 and 5) show MUSIC-type reconstructions of two anomalies without and with noise.

In Fig. 4, one sees clearly the presence of two anomalies. However, the one on the right, which is also deeper, is not as well rendered as the one on the left.

Bibliography and Open Questions

Thermal imaging of small anomalies has been considered in [17]. See also [29], where a realistic half-space model for thermal imaging was considered and accurate and robust reconstruction algorithms are designed.

It is worth mentioning that the inner expansions derived for the heat equation can be used to improve reconstruction in ultrasonic temperature imaging. The idea behind ultrasonic temperature imaging hinges on measuring local temperature near anomalies. The aim is to reconstruct anomalies with higher spatial and contrast resolution as compared to those obtained from boundary measurements alone. Further numerical investigations on this emerging topic are required.

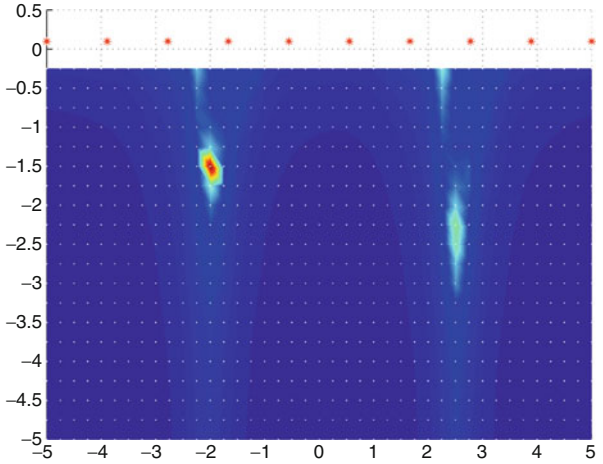


Fig. 4 Detection of anomalies using $n = 10$ heat sources equi-placed on the top

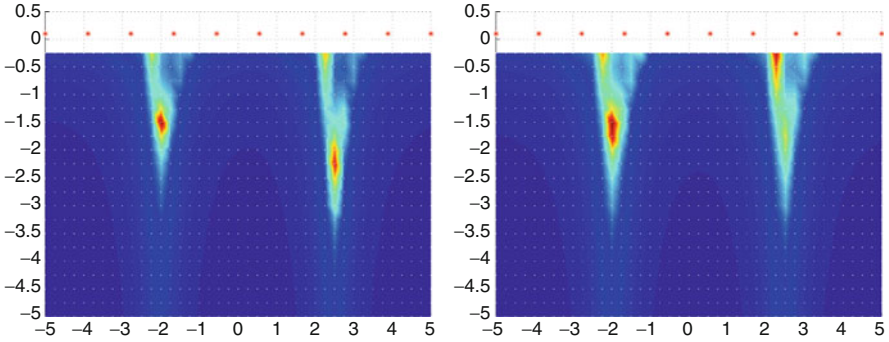


Fig. 5 Detection in the presence of 1 % (*on the left*) and 5 % (*on the right*) of measurement noise

5 Impediography

Physical Principles

Since all the present electrical impedance tomography technologies are only practically applicable in feature extraction of anomalies, improving electrical impedance tomography calls for innovative measurement techniques that incorporate structural information. A very promising direction of research is the recent magnetic resonance imaging technique, called current density imaging, which measures the

internal current density distribution. See the breakthrough work by Seo and his group described, for instance, in [65, 66, 89]. However, this technique has a number of disadvantages, among which the lack of portability and a potentially long imaging time. Moreover, it uses an expensive magnetic resonance imaging scanner.

Impediography is another mathematical direction for future electrical impedance tomography research in view of biomedical applications. It keeps the most important merits of electrical impedance tomography (real-time imaging, low cost, portability). It is based on the simultaneous measurement of an electric current and of acoustic vibrations induced by ultrasound waves. Its intrinsic resolution depends on the size of the focal spot of the acoustic perturbation, and thus, it may provide high-resolution images.

The core idea of impediography is to couple electric measurements to localized elastic perturbations. A body (a domain $\Omega \subset \mathbb{R}^2$) is electrically probed: one or several currents are imposed on the surface and the induced potentials are measured on the boundary. At the same time, a circular region of a few millimeters in the interior of Ω is mechanically excited by ultrasonic waves, which dilate this region. The measurements are made as the focus of the ultrasounds scans the entire domain. Several sets of measurements can be obtained by varying amplitudes of the ultrasound waves and the applied currents.

Within each disk of (small) volume, the conductivity is assumed to be constant per volume unit. At a point $x \in \Omega$, within a disk D of volume V_D, the electrical conductivity γ is defined in terms of a density ρ as $\gamma(x) = \rho(x)V_D$.

The ultrasonic waves induce a small elastic deformation of the disk D. If this deformation is isotropic, the material points of D occupy a volume V_D^p in the perturbed configuration, which at first order is equal to

$$V_D^p = V_D \left(1 + 2\frac{\Delta r}{r}\right),$$

where r is the radius of the disk D and Δr is the variation of the radius due to the elastic perturbation. As Δr is proportional to the amplitude of the ultrasonic wave, one obtains a proportional change of the deformation. Using two different ultrasonic waves with different amplitudes but with the same spot, it is therefore easy to compute the ratio V_D^p/V_D. As a consequence, the perturbed electrical conductivity γ^p satisfies

$$\forall\ x \in \Omega, \quad \gamma^p(x) = \rho(x)V_D^p = \gamma(x)\eta(x),$$

where $\eta(x) = V_D^p/V_D$ is a known function. One makes the following realistic assumptions: (1) the ultrasonic wave expands the zone it impacts and changes its conductivity, $\forall x \in \Omega$, $\eta(x) > 1$, and (2) the perturbation is not too small, $\eta(x) - 1 \gg V_D$.

Mathematical Model

Let u be the voltage potential induced by a current g, in the absence of ultrasonic perturbations. It is given by

$$\begin{cases} \nabla \cdot (\gamma(x)\nabla u) = 0 \text{ in } \Omega, \\ \gamma \dfrac{\partial u}{\partial \nu} = g \text{ on } \partial\Omega, \end{cases} \tag{44}$$

with the convention that $\int_{\partial\Omega} u = 0$. One supposes that the conductivity γ of the region close to the boundary of the domain is known, so that ultrasonic probing is limited to interior points. One denotes the region (open set) by Ω_1.

Let u_δ be the voltage potential induced by a current g, in the presence of ultrasonic perturbations localized in a disk-shaped domain $D := z + \delta B$ of volume $|D| = O\left(\delta^2\right)$. The voltage potential u_δ is a solution to

$$\begin{cases} \nabla \cdot (\gamma_\delta(x)\nabla u_\delta(x)) = 0 \text{ in } \Omega, \\ \gamma \dfrac{\partial u_\delta}{\partial \nu} = g \text{ on } \partial\Omega, \end{cases} \tag{45}$$

with the notation

$$\gamma_\delta(x) = \gamma(x)\left[1 + \chi(D)(x)\left(\eta(x) - 1\right)\right],$$

where $\chi(D)$ is the characteristic function of the domain D.

As the zone deformed by the ultrasound wave is small, one can view it as a small-volume perturbation of the background conductivity γ and seek an asymptotic expansion of the boundary values of $u_\delta - u$. The method of small-volume expansions shows that comparing u_δ and u on $\partial\Omega$ provides information about the conductivity. Indeed, one can prove that

$$\begin{aligned} \int_{\partial\Omega} (u_\delta - u) g \, d\sigma &= \int_D \gamma(x) \frac{(\eta(x) - 1)^2}{\eta(x) + 1} \nabla u \cdot \nabla u \, dx + o(|D|) \\ &= \gamma(z) \, |\nabla u(z)|^2 \int_D \frac{(\eta(x) - 1)^2}{\eta(x) + 1} \, dx + o(|D|). \end{aligned}$$

Note that because of assumption (2) at the end of the previous section, it follows that

$$\int_D \frac{(\eta(x) - 1)^2}{\eta(x) + 1} \, dx \geq C \, |D|$$

for some positive constant C. Therefore, one has

$$\gamma(z)\,|\nabla u(z)|^2 = \mathcal{E}(z) + o(1), \tag{46}$$

where the function $\mathcal{E}(z)$ is defined by

$$\mathcal{E}(z) = \left(\int_D \frac{(\eta(x)-1)^2}{\eta(x)+1}dx\right)^{-1} \int_{\partial\Omega} (u_\delta - u)g\,d\sigma. \tag{47}$$

By scanning the interior of the body with ultrasound waves, given an applied current g, one then obtains data from which one can compute the electrical energy

$$\mathcal{E}(z) := \gamma(z)|\nabla u(z)|^2$$

in an interior subregion of Ω. The new inverse problem is now to reconstruct γ, knowing $\mathcal{E}$.

Substitution Algorithm

The use of $\mathcal{E}$ leads one to transform (44), having two unknowns γ and u with highly nonlinear dependency on γ, into the following nonlinear PDE (the 0–Laplacian)

$$\begin{cases} \nabla_x \cdot \left(\dfrac{\mathcal{E}}{|\nabla u|^2}\nabla u\right) = 0 \text{ in } \Omega, \\ \dfrac{\mathcal{E}}{|\nabla u|^2}\dfrac{\partial u}{\partial \nu} = g \text{ on } \partial\Omega. \end{cases} \tag{48}$$

It is worth emphasizing that $\mathcal{E}$ is a known function, constructed from the measured data (47). Consequently, all the parameters entering in (48) are known. Thus, the ill-posed inverse problem in electrical impedance tomography is converted into a less-complicated direct problem (48).

The E-substitution algorithm, which will be explained below, uses two currents g_1 and g_2. One chooses this pair of current patterns to have $\nabla u_1 \times \nabla u_2 \neq 0$ for all $x \in \Omega$, where $u_i, i = 1, 2$, is the solution to (44). One refers to [66] and the references therein for an evidence of the possibility of such a choice. The E-substitution algorithm is based on an approximation of a linearized version of problem (48).

Suppose that γ is a small perturbation of conductivity profile γ_0 : $\gamma = \gamma_0 + \delta\gamma$. Let u_0 and $u = u_0 + \delta u$ denote the potentials corresponding to γ_0 and γ with the same Neumann boundary data g. It is easily seen that δu satisfies $\nabla \cdot (\gamma \nabla \delta u) = -\nabla \cdot (\delta\gamma \nabla u_0)$ in Ω with the homogeneous Dirichlet boundary condition. Moreover, from

$$\mathcal{E} = (\gamma_0 + \delta\gamma)|\nabla(u_0 + \delta u)|^2 \approx \gamma_0|\nabla u_0|^2 + \delta\gamma|\nabla u_0|^2 + 2\gamma_0\nabla u_0 \cdot \nabla\delta u,$$

after neglecting the terms $\delta\gamma\nabla u_0 \cdot \nabla\delta u$ and $\delta\gamma|\nabla\delta u|^2$, it follows that

$$\delta\gamma \approx \frac{\mathcal{E}}{|\nabla u_0|^2} - \gamma_0 - 2\gamma_0\frac{\nabla\delta u \cdot \nabla u_0}{|\nabla u_0|^2}.$$

The E-substitution algorithm is as follows. One starts from an initial guess for the conductivity γ and solves the corresponding Dirichlet conductivity problem

$$\begin{cases} \nabla \cdot (\gamma\nabla u_0) = 0 \text{ in } \Omega, \\ \qquad u_0 = \psi \text{ on } \partial\Omega. \end{cases}$$

The data ψ is the Dirichlet data measured as a response to the current g (say $g = g_1$) in the absence of elastic deformation. The discrepancy between the data and the guessed solution is

$$\epsilon_0 := \frac{\mathcal{E}}{|\nabla u_0|^2} - \gamma. \tag{49}$$

One then introduces a corrector, δu, computed as the solution to

$$\begin{cases} \nabla \cdot (\gamma\nabla\delta u) = -\nabla \cdot (\varepsilon_0\nabla u_0) \text{ in } \Omega, \\ \qquad\qquad \delta u = 0 \text{ on } \partial\Omega, \end{cases}$$

and updates the conductivity

$$\gamma := \frac{\mathcal{E} - 2\gamma\nabla\delta u \cdot \nabla u_0}{|\nabla u_0|^2}.$$

One iteratively updates the conductivity, alternating directions of currents (i.e., with $g = g_2$).

Consider a disk-shaped domain Ω, which contains three anomalies, an ellipse, an L-shaped domain, and a triangle. See Fig. 6.

Figure 7 shows the result of the reconstruction when measurements with very accurate precision for two directions of currents are available.

In the case of incomplete data, that is, if $\mathcal{E}$ is only known on a subset Ω' of the domain, one can follow an optimal control approach. See [39].

Bibliography and Open Questions

Impediography was proposed in [8], and the substitution algorithm proposed there. An optimal control approach for solving the inverse problem in impediography has

Fig. 6 Conductivity distribution

Fig. 7 Reconstruction test. From *left* to *right*, the initial guess, the collected data $\mathcal{E}$ for two directions of currents, and the reconstructed conductivity

been described in [39]. The inversion was considered as a minimization problem, and it was performed in two or three dimensions.

As pointed out in [71], the success of impediography depends on the feasibility of focusing ultrasound waves at an arbitrary point inside the body. Such a focusing, however, is quite tricky to achieve in practice. See, for instance, [85]. A method to extract the measurements corresponding to well-focused beams from the data obtained with unfocused waves has been proposed in [71].

An interesting problem is to study the sensitivity of the inversion methods to limitations on the intensities of the applied voltages, as electrical safety regulations limit the amount of the total current that patients can sustain. Another interesting problem is to reconstruct anisotropic conductivity distributions and to see whether or not impediography allows one to remove the obstruction to unique identifiability of the conductivity by electrical impedance tomography. In electrical impedance tomography, it is known that any change of variables of the background conductor that leaves the boundary fixed gives rise to a new anisotropic conductivity with the same boundary measurements [68].

6 Magneto-Acoustic Imaging

In magneto-acoustic imaging, a probe signal such as an acoustic wave or an electric current (or voltage) is applied to a biological tissue placed in a magnetic field. The probe signal produces, by the Lorentz force, an induced signal that is a function of the local electrical conductivity of the biological tissue. If the probe signal is an acoustic wave, then the induced signal is an electric current and the Lorentz force causes a local current density.

Induced boundary currents or pressure which are proportional to the local electrical conductivity can be measured to reconstruct the conductivity distribution with the spatial resolution of the ultrasound. The induced signal is detected, and an image of the local electrical conductivity of the specimen is generated based on the detected induced signal. The first method is referred as magneto-acousto-electrical tomography and the second one as magneto-acoustic tomography with magnetic induction.

Magneto-Acousto-Electrical Tomography

Physical Principles

In magneto-acousto-electrical imaging, an acoustic wave is applied to a biological tissue placed in a magnetic field. The probe signal produces by the Lorentz force an electric current that is a function of the local electrical conductivity of the biological tissue [80]. The mathematical basis for this magneto-acoustic imaging approach is provided, and an efficient algorithm for solving the inverse problem is proposed which is quite similar to the one designed for impediography.

Mathematical Model

Denote by $\gamma(x)$ the unknown conductivity, and let the voltage potential v be the solution to the conductivity problem

$$\begin{cases} \nabla \cdot \gamma \nabla v = 0 \text{ in } \Omega, \\ \qquad v = g \text{ on } \partial\Omega. \end{cases} \tag{50}$$

Suppose that the conductivity γ is a known constant on a neighborhood of the boundary $\partial\Omega$ and let γ_* denote $\gamma|_{\partial\Omega}$.

In magneto-acoustic imaging, ultrasonic waves are focused on regions of small diameter inside a body placed on a static magnetic field. The oscillation of each small region results in frictional forces being applied to the ions, making them move. In the presence of a magnetic field, the ions experience a Lorentz force. This

gives rise to a localized current density within the medium. The current density is proportional to the local electrical conductivity [80]. In practice, the ultrasounds impact a spherical or ellipsoidal zone, of a few millimeters in diameter. The induced current density should thus be sensitive to conductivity variations at the millimeter scale, which is the precision required for breast cancer diagnostics.

Let $z \in \Omega$ and D be a small impact zone around the point z. The created current by the Lorentz force density is given by

$$\mathbf{J}_z(x) = c\chi(D)(x)\gamma(x)\mathbf{e}, \tag{51}$$

for some constant c and a constant unit vector $\mathbf{e}$, both of which are independent of z. With the induced current $\mathbf{J}_z$, the new voltage potential, denoted by u_z, satisfies

$$\begin{cases} \nabla \cdot (\gamma \nabla u_z + \mathbf{J}_z) = 0 \text{ in } \Omega, \\ u_z = g \text{ on } \partial\Omega. \end{cases}$$

According to (51), the induced electrical potential $w_z := v - u_z$ satisfies the conductivity equation:

$$\begin{cases} \nabla \cdot \gamma \nabla w_z = c\nabla \cdot (\chi(D)\gamma\mathbf{e}) \text{ for } x \in \Omega, \\ w_z(x) = 0 \text{ for } x \in \partial\Omega. \end{cases} \tag{52}$$

The inverse problem for the magneto-acousto-electrical imaging is to reconstruct the conductivity profile γ from boundary measurements of $\frac{\partial u_z}{\partial \nu}|_{\partial\Omega}$ or equivalently $\frac{\partial w_z}{\partial \nu}|_{\partial\Omega}$ for $z \in \Omega$.

Substitution Algorithm

Since γ is assumed to be constant in D and $|D|$ is small, one obtains using Green's identity

$$\int_{\partial\Omega} \gamma_* \frac{\partial w_z}{\partial \nu} g d\sigma \approx -c|D|\nabla(\gamma v)(z) \cdot \mathbf{e}\,. \tag{53}$$

The relation (53) shows that, by scanning the interior of the body with ultrasound waves, $c\nabla(\gamma v)(z) \cdot \mathbf{e}$ can be computed from the boundary measurements $\frac{\partial w_z}{\partial \nu}|_{\partial\Omega}$ in Ω. If one can rotate the subject, then $c\nabla(\gamma v)(z)$ for any z in Ω can be reconstructed. In practice, the constant c is not known. But, since γv and $\partial(\gamma v)/\partial \nu$ on the boundary of Ω are known, one can recover c and γv from $c\nabla(\gamma v)$ in a constructive way [11].

The new inverse problem is now to reconstruct the contrast profile γ, knowing

$$\mathcal{E}(z) := \gamma(z)v(z) \tag{54}$$

for a given boundary potential g, where v is the solution to (50).

In view of (54), v satisfies

$$\begin{cases} \nabla \cdot \dfrac{\mathcal{E}}{v} \nabla v = 0 \text{ in } \Omega, \\ \qquad v = g \text{ on } \partial\Omega. \end{cases} \tag{55}$$

If one solves (55) for v, then (54) yields the conductivity contrast γ. Note that to be able to solve (55), one needs to know the coefficient $\mathcal{E}(z)$ for all z, which amounts to scanning all the points $z \in \Omega$ by the ultrasonic beam.

Observe that solving (55) is quite easy mathematically: if one puts $w = \ln v$, then w is the solution to

$$\begin{cases} \nabla \cdot \mathcal{E} \nabla w = 0 \text{ in } \Omega, \\ \qquad w = \ln g \text{ on } \partial\Omega, \end{cases} \tag{56}$$

as long as $g > 0$. Thus, if one solves (56) for w, then $v := e^w$ is the solution to (55). However, taking an exponent may amplify the error which already exists in the computed data $\mathcal{E}$. In order to avoid this numerical instability, one solves (55) iteratively. To do so, one can adopt an iterative scheme similar to the one proposed in the previous section.

Start with γ_0 and let v_0 be the solution of

$$\begin{cases} \nabla \cdot \gamma_0 \nabla v_0 = 0 \text{ in } \Omega, \\ \qquad v_0 = g \text{ on } \partial\Omega. \end{cases} \tag{57}$$

According to (54), the updates, $\gamma_0 + \delta\gamma$ and $v_0 + \delta v$, should satisfy

$$\gamma_0 + \delta\gamma = \frac{\mathcal{E}}{v_0 + \delta v}, \tag{58}$$

where

$$\begin{cases} \nabla \cdot (\gamma_0 + \delta\gamma) \nabla (v_0 + \delta v) = 0 \text{ in } \Omega, \\ \qquad\qquad\qquad\qquad \delta v = 0 \text{ on } \partial\Omega, \end{cases}$$

or equivalently

$$\begin{cases} \nabla \cdot \gamma_0 \nabla \delta v + \nabla \cdot \delta\gamma \nabla v_0 = 0 \text{ in } \Omega, \\ \qquad\qquad\qquad\qquad \delta v = 0 \text{ on } \partial\Omega. \end{cases} \tag{59}$$

One then linearizes (58) to have

$$\gamma_0 + \delta\gamma = \frac{\mathcal{E}}{v_0(1 + \delta v/v_0)} \approx \frac{\mathcal{E}}{v_0}\left(1 - \frac{\delta v}{v_0}\right). \tag{60}$$

Fig. 8 Reconstruction test. From *left* to *right*, the conductivity distribution, the initial guess, and the reconstructed conductivity after three iterations

Thus,

$$\delta\gamma = -\frac{\mathcal{E}\delta v}{v_0^2} - \delta, \quad \delta = -\frac{\mathcal{E}}{v_0} + \gamma_0. \tag{61}$$

One then finds δv by solving

$$\begin{cases} \nabla \cdot \gamma_0 \nabla \delta v - \nabla \cdot \left(\frac{\mathcal{E}\nabla v_0}{v_0^2} \delta v \right) = \nabla \cdot \delta \nabla v_0 \text{ in } \Omega, \\ \qquad\qquad\qquad\qquad\qquad\qquad \delta v = 0 \text{ on } \partial\Omega. \end{cases} \tag{62}$$

Figure 8 shows the result of the reconstruction when very accurate measurements for two Dirichlet boundary conditions, $g = g_1, g_2$, are available.

In the case of incomplete data, that is, if $\mathcal{E}$ is only known on a subset ω of the domain, one can follow an optimal control approach. See [11].

Magneto-Acoustic Imaging with Magnetic Induction

Physical Principles

In the magneto-acoustic tomography with magnetic induction, pulsed magnetic stimulation by the ultrasound beam is imposed on an object placed in a static magnetic field. The magnetic stimulation can be considered as an ideal pulsed distribution over time. The magnetically induced eddy current is then subject to a Lorentz force. This in turn creates a pressure wave that can be detected using an ultrasound hydrophone [80]. The magneto-acoustic tomography with magnetic induction uses this acoustic pressure wave to reconstruct the conductivity distribution of the sample as the focus of the ultrasound beam scans the entire domain.

Mathematical Model

Let γ be the conductivity distribution of the object as before. Denoting the constant magnetic field as B_0 and the magnetically induced current density distribution as $\mathbf{J}_z(x)$ with z indicating the location of the magnetic stimulation, the Lorentz force is given by

$$\mathbf{J}_z(x) \times B_0 \delta_{t=0} = c\,\chi(D)(x)\gamma(x)\mathbf{e}\delta_{t=0},$$

where D is the impact zone which is a small neighborhood of z as before, and c is a constant independent of z and x. Then the wave equation governing the pressure distribution p_z can be written as

$$\frac{\partial^2 p_z}{\partial t^2} - c_s^2 \Delta p_z = 0, \quad x \in \Omega, \quad t \in]0, T[, \tag{63}$$

for some final observation time T, where c_s is the acoustic speed in Ω. The pressure satisfies the Dirichlet boundary condition

$$p_z = 0 \quad \text{on } \partial\Omega \times]0, T[\tag{64}$$

and the initial conditions

$$p_z|_{t=0} = 0 \quad \text{and} \quad \left.\frac{\partial p_z}{\partial t}\right|_{t=0} = -c\nabla \cdot (\chi(D)\gamma\mathbf{e}) \quad \text{in } \Omega. \tag{65}$$

The inverse problem for the magneto-acoustic tomography with magnetic induction is to determine the conductivity distribution γ in Ω from boundary measurements of $\frac{\partial p_z}{\partial \nu}$ on $\partial\Omega \times]0, T[$ for all $z \in \Omega$. Suppose that T is large enough so that

$$T > \frac{\text{diam}(\Omega)}{c_s}, \tag{66}$$

which says that the observation time is long enough for the wave initiated at z to reach the boundary $\partial\Omega$.

Reconstruction Algorithm

The algorithms for the magneto-acoustic tomography with magnetic induction available in the literature are limited to unbounded media. They use the spherical Radon transform inversion. However, the pressure field is significantly affected by the acoustic boundary conditions at the tissue–air interface, where the pressure must vanish. Thus, one cannot base magneto-acoustic imaging on pressure measurements made over a free surface. Instead, one can use the following algorithm.

Let w satisfy

$$\frac{\partial^2 w}{\partial t^2} - c_s^2 \Delta w = 0 \quad \text{in } \Omega \times]0, T[, \tag{67}$$

with the final conditions

$$w|_{t=T} = \frac{\partial w}{\partial t}\Big|_{t=T} = 0 \quad \text{in } \Omega. \tag{68}$$

Since γ is constant on D, one can prove that the following identity holds:

$$\int_0^T \int_{\partial\Omega} \frac{\partial p_z}{\partial \nu}(x,t) w(x,t)\, d\sigma(x)\, dt = \frac{c}{c_s^2}\gamma(z) \int_D \mathbf{e} \cdot \nabla w(x,0) dx. \tag{69}$$

Suppose that $d = 3$. For $y \in \mathbb{R}^3 \backslash \overline{\Omega}$, define the probe function

$$w_y(x,t) := \frac{\delta\left(t + \tau - \frac{|x-y|}{c_s}\right)}{4\pi|x-y|} \quad \text{in } \Omega \times]0, T[, \tag{70}$$

where $\tau := \frac{|y-z|}{c_s}$. The function w_y is a Green's function corresponding to retarded potentials. Choosing w_y as a test function in (69) yields the new identity

$$c\gamma(z) = \frac{c_s^2}{\int_D \mathbf{e} \cdot \nabla w_y(x,0) dx} \int_0^T \int_{\partial\Omega} \frac{\partial p_z}{\partial \nu}(x,t) w_y(x,t)\, d\sigma(x)\, dt. \tag{71}$$

The quantity $\int_D \mathbf{e} \cdot \nabla w_y(x,0) dx$ can be explicitly computed. In particular, if the source point y is such that $z - y$ is parallel to $\mathbf{e}$ and D is a sphere of radius r (and center z), then

$$c\gamma(z) = -\frac{c_s}{\frac{r^2}{2|z-y|^2} - \frac{r^4}{4|z-y|^4}} \int_0^T \int_{\partial\Omega} \frac{\partial p_z}{\partial \nu}(x,t) w_y(x,t)\, d\sigma(x)\, dt, \tag{72}$$

provided that γ is constant on D. But since r is sufficiently small, one obtains

$$c\gamma(z) \approx -\frac{2c_s|z-y|^2}{r^2} \int_0^T \int_{\partial\Omega} \frac{\partial p_z}{\partial \nu}(x,t) w_y(x,t)\, d\sigma(x)\, dt. \tag{73}$$

Formula (73) can be used to effectively compute the conductivity contrast in Ω with a resolution of order the size of the ultrasound beam. It is worth emphasizing that unlike magneto-acousto-electrical imaging, in magneto-acoustic tomography with magnetic induction, it suffices to excite the local spot at z in order to obtain the value $c\gamma(z)$, as clearly shown by (73).

Bibliography and Open Questions

The feasibility of magneto-acoustic imaging has been demonstrated in [54, 75, 76]. The mathematical and numerical modeling described in this section is from [11]. As it will be shown in Sect. 8, the approach for the magneto-acoustic tomography with magnetic induction can be used in photo-acoustic imaging.

It would be interesting to prove the convergence of the proposed iterative scheme for magneto-acousto-electrical tomography. Another important problem is to design an efficient inversion algorithm for magneto-acoustic tomography with magnetic induction when the acoustic speed fluctuates randomly.

7 Magnetic Resonance Elastography

Physical Principles

Extensive work has been carried out in the past decade to image, by inducing motion, the elastic properties of human soft tissues. This wide application field, called elasticity imaging or elastography, is based on the initial idea that shear elasticity can be correlated with the pathological state of tissues. Several techniques arose according to the type of mechanical excitation chosen (static compression, monochromatic, or transient vibration) and the way these excitations are generated (externally or internally). Different imaging modalities can be used to estimate the resulting tissue displacements.

Magnetic resonance elastography (MRE) is a new way of realizing the idea of elastography. It can directly visualize and quantitatively measure the displacement field in tissues subject to harmonic mechanical excitation at low frequencies. A phase-contrast magnetic resonance imaging technique is used to spatially map and measure the complete three-dimensional displacement patterns. From this data, local quantitative values of shear modulus can be calculated, and images that depict tissue elasticity or stiffness can be generated. The inverse problem for magnetic resonance elastography is to determine the shape and the elastic parameters of an elastic anomaly from internal measurements of the displacement field. In most cases, the most significant elastic parameter is the stiffness coefficient.

In biological media, the compression modulus is four to six orders higher than the shear modulus. One can prove that, as the compression modulus goes to $+\infty$, the Lamé system converges to the modified Stokes system. By reducing the elasticity system to a modified Stokes system, one removes the compression modulus from consideration.

Mathematical Model

Consider the modified Stokes system, i.e., the problem of determining $\mathbf{v}$ and q in a domain Ω from the conditions:

$$\begin{cases} \left(\Delta + \kappa^2\right) \mathbf{v} - \nabla q = 0, \\ \nabla \cdot \mathbf{v} = 0\,, \\ \mathbf{v}|_{\partial\Omega} = \mathbf{g}\,. \end{cases} \tag{74}$$

Problem (74) governs elastic wave propagation in nearly incompressible homogeneous media.

Let $(G_{il})_{i,l=1}^{d}$ be the Dirichlet Green function for the operator in (74), i.e., for $y \in \Omega$,

$$\begin{cases} \left(\Delta_x + \kappa^2\right) G_{il}(x,y) - \dfrac{\partial F_i(x-y)}{\partial x_l} = \delta_{il}\delta_y(x) \quad \text{in } \Omega\,, \\ \displaystyle\sum_{l=1}^{d} \frac{\partial}{\partial x_l} G_{il}(x,y) = 0 \quad \text{in } \Omega\,, \\ G_{il}(x,y) = 0 \quad \text{on } \partial\Omega\,. \end{cases} \tag{75}$$

Denote by $(\mathbf{e}_1, \ldots, \mathbf{e}_d)$ an orthonormal basis of $\mathbb{R}^d$. Let $d(\xi) := (1/d) \sum_k \xi_k \mathbf{e}_k$ and $\hat{\mathbf{v}}_{pq}$, for $p, q = 1, \ldots, d$, be the solution to

$$\begin{cases} \mu \Delta \hat{\mathbf{v}}_{pq} + \nabla \hat{p} = 0 \quad \text{in } \mathbb{R}^d \backslash \overline{B}, \\ \tilde{\mu} \Delta \hat{\mathbf{v}}_{pq} + \nabla \hat{p} = 0 \quad \text{in } B, \\ \hat{\mathbf{v}}_{pq}|_- - \hat{\mathbf{v}}_{pq}|_+ = 0 \quad \text{on } \partial B, \\ \left(\hat{p}\mathbf{N} + \tilde{\mu} \dfrac{\partial \hat{\mathbf{v}}_{pq}}{\partial \mathbf{N}} \right)|_- - \left(\hat{p}\mathbf{N} + \mu \dfrac{\partial \hat{\mathbf{v}}_{pq}}{\partial \mathbf{N}} \right)|_+ = 0 \quad \text{on } \partial B, \\ \nabla \cdot \hat{\mathbf{v}}_{pq} = 0 \quad \text{in } \mathbb{R}^d, \\ \hat{\mathbf{v}}_{pq}(\xi) \to \xi_p \mathbf{e}_q - \delta_{pq} d(\xi) \quad \text{as } |\xi| \to \infty, \\ \hat{p}(\xi) \to 0 \quad \text{as } |\xi| \to +\infty. \end{cases} \tag{76}$$

Here, $\partial \mathbf{v}/\partial \mathbf{N} = (\nabla \mathbf{v} + (\nabla \mathbf{v})^*) \cdot \mathbf{N}$ and $(\nabla \mathbf{v})^*$ denotes the transpose of the matrix $\nabla \mathbf{v}$.

Define the viscous moment tensor $(V_{ijpq})_{i,j,p,q=1,\ldots,d}$ by

$$V_{ijpq} := (\tilde{\mu} - \mu) \int_B \nabla \hat{\mathbf{v}}_{pq} \cdot (\nabla(\xi_i \mathbf{e}_j) + \nabla(\xi_i \mathbf{e}_j)^*)\, d\xi. \tag{77}$$

Consider an elastic anomaly D inside a nearly compressible medium Ω. The anomaly D has a shear modulus $\tilde{\mu}$ different from that of Ω, μ. The displacement field $\mathbf{u}$ solves the following transmission problem for the modified Stokes problem:

$$\begin{cases} \left(\mu\Delta + \omega^2\right)\mathbf{u} + \nabla p = 0 \quad \text{in } \Omega\backslash\overline{D}, \\ \left(\tilde{\mu}\Delta + \omega^2\right)\mathbf{u} + \nabla p = 0 \quad \text{in } D, \\ \mathbf{u}\big|_- = \mathbf{u}\big|_+ \quad \text{on } \partial D, \\ (p|_+ - p|_-)\mathbf{N} + \mu\dfrac{\partial \mathbf{u}}{\partial \mathbf{N}}\bigg|_+ - \tilde{\mu}\dfrac{\partial \mathbf{u}}{\partial \mathbf{N}}\bigg|_- = 0 \quad \text{on } \partial D, \\ \nabla\cdot\mathbf{u} = 0 \quad \text{in } \Omega, \\ \mathbf{u} = \mathbf{g} \quad \text{on } \partial\Omega, \\ \displaystyle\int_\Omega p = 0, \end{cases} \tag{78}$$

where $\mathbf{g} \in L^2(\partial\Omega)$ satisfies the compatibility condition $\displaystyle\int_{\partial\Omega} \mathbf{g}\cdot\mathbf{N} = 0$.

The inverse problem consists of reconstructing $\tilde{\mu}$ and the shape of the inclusion D from internal measurements of $\mathbf{u}$.

Asymptotic Analysis of Displacement Fields

Let $(\mathbf{U}, q)$ denote the background solution to the modified Stokes system in the absence of any anomalies, that is, the solution to

$$\begin{cases} \left(\mu\Delta + \omega^2\right)\mathbf{U} + \nabla q = 0 \quad \text{in } \Omega, \\ \nabla\cdot\mathbf{U} = 0 \quad \text{in } \Omega, \\ \mathbf{U} = \mathbf{g} \quad \text{on } \partial\Omega, \\ \displaystyle\int_\Omega q = 0. \end{cases} \tag{79}$$

The following asymptotic expansions hold.

Theorem 8 (Expansions of the displacement field). *Suppose that $D = \delta B + z$, and let u be the solution of (78), where $0 < \tilde{\mu} \neq \mu < +\infty$.*

(i) *The following inner expansion holds:*

$$\mathbf{u}(x) \approx \mathbf{U}(z) + \delta \sum_{p,q=1}^{d} \partial_q \mathbf{U}(z)_p \hat{\mathbf{v}}_{pq}\left(\frac{x-z}{\delta}\right) \quad \text{for } x \text{ near } z, \tag{80}$$

where $\hat{\mathbf{v}}_{pq}$ is defined by (76).

(ii) *Let (V_{ijpq}) be the viscous moment tensor defined by (77). The following outer expansion holds uniformly for $x \in \partial\Omega$:*

$$(\mathbf{u}-\mathbf{U})(x) \approx \delta^d \left[\sum_{i,j,p,q,\ell=1}^{d} \mathbf{e}_\ell \partial_j G_{\ell i}(x,z) \partial_q \mathbf{U}(z)_p V_{ijpq} \right], \tag{81}$$

where V_{ijpq} is given by (77), and the Green function $(G_{il})_{i,l=1}^d$ is defined by (75) with $\kappa^2 = \omega^2/\mu$, μ being the shear modulus of the background medium.

The notion of a viscous moment tensor extends the notion of a polarization tensor to quasi-incompressible elasticity. The viscous moment tensor, V, characterizes all the information about the elastic anomaly that can be learned from the leading-order term of the outer expansion (81). It can be explicitly computed for disks and ellipses in the plane and balls and ellipsoids in three-dimensional space. If B is a two-dimensional disk, then

$$V = 4\,|B|\,\mu \frac{(\tilde{\mu}-\mu)}{\tilde{\mu}+\mu} P,$$

where $P = \left(P_{ijpq}\right)$ is the orthogonal projection from the space of symmetric matrices onto the space of symmetric matrices of trace zero, i.e.,

$$P_{ijpq} = \frac{1}{2}\left(\delta_{ip}\delta_{jq} + \delta_{iq}\delta_{jp}\right) - \frac{1}{d}\delta_{ij}\delta_{pq}.$$

If B is an ellipse of the form

$$\frac{x_1^2}{a^2} + \frac{x_2^2}{b^2} = 1, \quad a \geq b > 0, \tag{82}$$

then the viscous moment tensor for B is given by

$$\begin{cases} V_{1111} = V_{2222} = -V_{1122} = -V_{2211} = |B| \dfrac{2\mu(\tilde{\mu}-\mu)}{\mu+\tilde{\mu}-(\tilde{\mu}-\mu)m^2}, \\ V_{1212} = V_{2112} = V_{1221} = V_{2121} = |B| \dfrac{2\mu(\tilde{\mu}-\mu)}{\mu+\tilde{\mu}+(\tilde{\mu}-\mu)m^2}, \\ \text{the remaining terms are zero,} \end{cases} \tag{83}$$

where $m = (a-b)/(a+b)$.

If B is a ball in three dimensions, the viscous moment tensor associated with B and an arbitrary $\tilde{\mu}$ is given by

$$\begin{cases} V_{iiii} = \dfrac{20\mu|B|}{3}\dfrac{\tilde{\mu}-\mu}{2\tilde{\mu}+3\mu}, \quad V_{iijj} = -\dfrac{10\mu|B|}{3}\dfrac{\tilde{\mu}-\mu}{2\tilde{\mu}+3\mu}\ (i \neq j), \\ V_{ijij} = V_{ijji} = 5\mu|B|\dfrac{\tilde{\mu}-\mu}{2\tilde{\mu}+3\mu},\ (i \neq j), \\ \text{the remaining terms are zero.} \end{cases} \tag{84}$$

Theorem 9 (Properties of the viscous moment tensor). *For $0 < \tilde{\mu} \neq \mu < +\infty$, let $V = (V_{ijpq})_{i,p,q=1}^{d}$ be the viscous moment tensor associated with the bounded domain B in $\mathbb{R}^d$ and the pair of shear modulus $(\tilde{\mu}, \mu)$. Then*

(i) *For $i, j, p, q = 1, \ldots, d$,*

$$V_{ijpq} = V_{jipq}, \quad V_{ijpq} = V_{ijqp}, \quad V_{ijpq} = V_{pqij}. \tag{85}$$

(ii) *One has*

$$\sum_p V_{ijpp} = 0 \quad \text{for all } i, j \quad \text{and} \quad \sum_i V_{iipq} = 0 \quad \text{for all } p, q,$$

or equivalently, $V = PVP$.

(iii) *The tensor V is positive (negative, resp.) definite on the space of symmetric matrices of trace zero if $\tilde{\mu} > \mu$ ($\tilde{\mu} < \mu$, resp.).*

(iv) *The tensor $(1/(2\mu))\,V$ satisfies the following bounds:*

$$\mathrm{Tr}\left(\frac{1}{2\mu}V\right) \leq |B|\left(\frac{\tilde{\mu}}{\mu}-1\right)\left((d-1)\frac{\mu}{\tilde{\mu}} + \frac{d(d-1)}{2}\right), \tag{86}$$

$$\mathrm{Tr}\left(\frac{1}{2\mu}V\right)^{-1} \leq \frac{1}{|B|\left(\frac{\tilde{\mu}}{\mu}-1\right)}\left((d-1)\frac{\tilde{\mu}}{\mu} + \frac{d(d-1)}{2}\right), \tag{87}$$

where for $C = (C_{ijpq})$, $\mathrm{Tr}(C) := \sum_{i,j=1}^{d} C_{ijij}$.

Note that the viscous moment tensor, V, is a four tensor and can be regarded, because of its symmetry, as a linear transformation on the space of symmetric matrices. Note also that, in view of Theorem 2, the right-hand sides of (86) and (87) are exactly in the two-dimensional case ($d = 2$) the Hashin–Shtrikman bounds (9) for the polarization tensor associated with the same domain B and the conductivity contrast $k = \tilde{\mu}/\mu$.

Numerical Methods

Let $\mathbf{u}$ be the solution to the modified Stokes system (78). The inverse problem in the magnetic resonance elastography is to reconstruct the shape and the shear modulus of the anomaly D from internal measurements of $\mathbf{u}$.

Based on the inner asymptotic expansion (80) of $\delta\mathbf{u}$ $(:= \mathbf{u} - \mathbf{U})$ of the perturbations in the displacement field that are due to the presence of the anomaly, a reconstruction method of binary level set type can be designed.

The first step for the reconstruction procedure is to locate the anomaly. This can be done using the outer expansion of $\delta\mathbf{u}$, i.e., an expansion far away from the elastic anomaly.

Suppose that z is reconstructed. Since the representation $D = z + \delta B$ is not unique, one can fix δ. One uses a binary level set representation f of the scaled domain B:

$$f(x) = \begin{cases} 1, \, x \in B \, , \\ -1, \, x \in \mathbb{R}^3 \backslash \overline{B}. \end{cases} \tag{88}$$

Let

$$2h(x) = \tilde{\mu}\left(f\left(\frac{x-z}{\delta}\right) + 1\right) - \mu\left(f\left(\frac{x-z}{\delta}\right) - 1\right) \tag{89}$$

and let β be a regularization parameter. Then the second step is to fix a window W (containing z) and solve the following constrained minimization problem

$$\begin{aligned} \min_{\tilde{\mu}, f} L(f, \tilde{\mu}) = & \frac{1}{2} \left\| \delta\mathbf{u}(x) - \delta \sum_{p,q=1}^{d} \partial_q \mathbf{U}(z)_p \hat{\mathbf{v}}_{pq}\left(\frac{x-z}{\delta}\right) + \nabla\mathbf{U}(z)(x-z) \right\|^2_{L^2(W)} \\ & + \beta \int_W |\nabla h(x)| \, dx, \end{aligned} \tag{90}$$

subject to (76). Here, $\int_W |\nabla h| \, dx$ is the total variation of the shear modulus and $|\nabla h|$ is understood as a measure:

$$\int_W |\nabla h| = \sup \left\{ \int_W h \nabla \cdot \mathbf{v} \, dx, \;\; \mathbf{v} \in C_0^1(W) \text{ and } |\mathbf{v}| \leq 1 \text{ in } W \right\}.$$

This regularization indirectly controls both the length of the level curves and the jumps in the coefficients.

The local character of the method is due to the decay of

$$\delta \sum_{p,q=1}^{d} \partial_q \mathbf{U}(z)_p \hat{\mathbf{v}}_{pq}\left(\frac{\cdot - z}{\delta}\right) - \nabla\mathbf{U}(z)(\cdot - z)$$

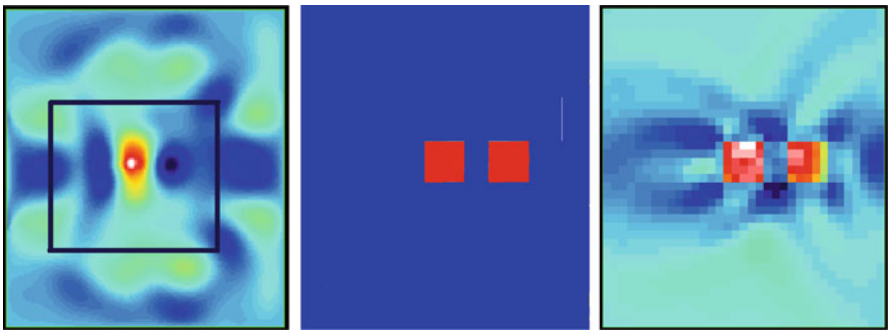


Fig. 9 Reconstruction using the data on the whole domain on the *left*, a zoom on the anomaly in the *middle*, and on the *right* the reconstruction limited on the subregion defined by the boxed region on the *left*

away from z. This is one of the main features of the method. In the presence of noise, because of a trade-off between accuracy and stability, one has to choose carefully the size of W. As it has been shown in [13], the size of W should not be too small in order to preserve some stability and not too big so that one can gain some accuracy. See Fig. 9.

The minimization problem (90) corresponds to a minimization with respect to $\tilde{\mu}$ followed by a step of minimization with respect to f. The minimization steps are over the set of $\tilde{\mu}$ and f and can be performed using a gradient-based method with a line search. Of importance are the optimal bounds satisfied by the viscous moment tensor V. One should check at each step whether the bounds (86) and (87) on V are satisfied or not. In the case where they are not, one has to restate the value of $\tilde{\mu}$. Another way to deal with (86) and (87) is to introduce them into the minimization problem (90) as a constraint. Set $\alpha = \mathrm{Tr}(V)$ and $\beta = \mathrm{Tr}\left(V^{-1}\right)$ and suppose for simplicity that $\tilde{\mu} > \mu$. Then, (86) and (87) can be rewritten (when $d = 3$) as follows:

$$\begin{cases} \alpha \leq 2(\tilde{\mu} - \mu)\left(3 + \dfrac{2\mu}{\tilde{\mu}}\right)|D|, \\ \dfrac{2\mu(\tilde{\mu} - \mu)}{3\mu + 2\tilde{\mu}}|D| \leq \beta^{-1}. \end{cases} \tag{91}$$

Bibliography and Open Questions

Magnetic resonance elastography was first proposed in [81]. The results provided on this technique are from [14]. Theorem 8 and the results on the viscous moment tensor in Theorem 9 are from [15].

In general, the elastic parameters of biological tissues show anisotropic properties, that is, the local value of elasticity is different in the different spatial directions [91] and also viscous properties. It would be very interesting to extend the algorithm described in this section for detecting the shape of an elastic anomaly and the viscosity and the anisotropy in its shear modulus. The study of the dependence of the shear modulus as a function of the frequency is also important [90].

8 Photo-Acoustic Imaging of Small Absorbers

Physical Principles

In photo-acoustic imaging, optical energy absorption causes thermoelastic expansion of the tissue, which in turn leads to propagation of a pressure wave. This signal is measured by transducers distributed on the boundary of the object, which is in turn used for imaging optical properties of the object. The significance of photo-acoustic imaging is to provide images of optical contrasts (based on the optical absorption) with the resolution of ultrasound.

In pure optical imaging, optical scattering in soft tissues degrades spatial resolution significantly with depth. As for electrical impedance tomography, even though pure optical imaging is very sensitive to optical absorption, it can only provide a spatial resolution of the order of 1 cm at centimeter depths. As discussed before, pure conventional ultrasound imaging is based on the detection of the mechanical properties (acoustic impedance) in biological soft tissues. It can provide good spatial resolution because of its millimetric wavelength and weak scattering at megahertz frequencies.

If the medium is acoustically homogeneous and has the same acoustic properties as the free space, then the boundary of the object plays no role and the optical properties of the medium can be extracted from the measurements of the pressure wave by inverting a spherical Radon transform.

In the more realistic situation, where a boundary condition has to be imposed on the pressure field, such an inversion formula does not hold. Using asymptotic analysis, one can develop an efficient approach for reconstructing absorbing regions and absorbing energy density inside a bounded domain from boundary data. One can also reconstruct the optical absorption coefficient. In general, it is not possible to infer physiological parameters from the absorbing energy density. It is the optical absorption coefficient distribution that directly correlates with tissue structural and functional information such as blood oxygenation.

Mathematical Model

Let $D_l, l = 1, \ldots, m$, be m absorbing domains inside the nonabsorbing background-bounded medium $\Omega \subset \mathbb{R}^d, d = 2$ or 3. In an acoustically homogeneous medium, the photo-acoustic effect is described by the following equation:

$$\frac{\partial^2 p}{\partial t^2}(x,t) - c_s^2 \Delta p(x,t) = \gamma \frac{\partial H}{\partial t}(x,t), \quad x \in \Omega, \quad t \in \mathbb{R}, \tag{92}$$

where c_s is the acoustic speed in Ω, γ the dimensionless Grüneisen coefficient in Ω, and $H(x,t)$ a heat source function (absorbed energy per unit time per unit volume).

Assuming the stress-confinement condition, the source term can be modeled as $\gamma H(x,t) = \delta(t)A(x)$, where the absorbed optical energy density times the Grüneisen coefficient $A = \sum_{l=1}^{m} A_l \chi(D_l)$ and A_l are constants. Under this assumption, the pressure in an acoustically homogeneous medium obeys the following wave equation:

$$\frac{\partial^2 p}{\partial t^2}(x,t) - c_s^2 \Delta p(x,t) = 0, \quad x \in \Omega, \quad t \in]0,T[,$$

for some final observation time T. The pressure satisfies the Dirichlet boundary condition

$$p = 0 \quad \text{on } \partial\Omega \times]0,T[$$

and the initial conditions

$$p|_{t=0} = \sum_{l=1}^{m} \chi(D_l) A_l \quad \text{and} \quad \left.\frac{\partial p}{\partial t}\right|_{t=0} = 0 \quad \text{in } \Omega.$$

Suppose that T satisfies (66). The inverse problem in photo-acoustic imaging is to determine the supports of nonzero optical absorption $(D_l, l = 1, \ldots, m)$ in Ω and $A(x)$ from boundary measurements of $\frac{\partial p}{\partial \nu}$ on $\partial\Omega \times]0,T[$.

Reconstruction Algorithms

Analogously to (71), the following identity holds:

$$\frac{1}{c_s^2} \sum_{l=1}^{m} A_l \int_D \partial_t w_y(x,0;\tau) dx = \int_0^T \int_{\partial\Omega} \frac{\partial p}{\partial \nu}(x,t) w_y(x,t;\tau)\, d\sigma(x)\, dt, \tag{93}$$

where the probe function w_y is given by (70).

Determination of Location

Suppose for simplicity that there is only one absorbing object ($m = 1$) which is denoted by D ($= z + \delta B$). Identity (93) shows that the imaging functional

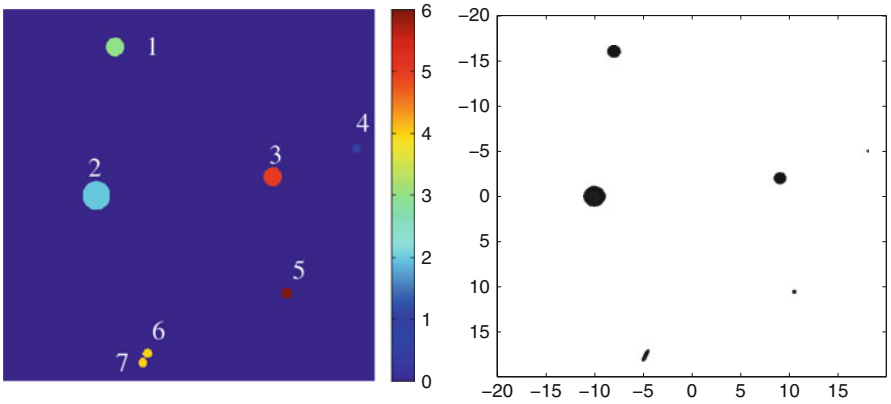


Fig. 10 Real configuration of the medium on the *left* – there are seven optical anomalies of various size and absorption. Reconstructed configuration on the *right* – anomalies 6 and 7 are reconstructed as a single anomaly

$$W(\tau, y) := \int_0^T \int_{\partial\Omega} \frac{\partial p}{\partial \nu}(x,t) w_y(x,t;\tau)\, d\sigma(x)\, dt \tag{94}$$

is nonzero only on the interval $]\tau_a, \tau_e[$, where $\tau_a = \mathrm{dist}(y, D)/c_s$ is the first τ for which the sphere of center y and radius τ hits D and τ_e is the last τ for which such sphere hits D. This gives a simple way to detect the location (by changing the source point y and taking intersection of spheres). The functional $W(\tau, y)$ can be used to probe the medium as a function of τ and y. For fixed y, it is a one-dimensional function and is related to time reversal in the sense that it is a convolution with a reversed wave.

A result of numerical simulation to validate the location search algorithm is given in Fig. 10.

Estimation of Absorbing Energy

Consider first the three-dimensional case. If D is a sphere with $A(x) = A\chi(D)$, then one has

$$\delta^2 A \approx c_s |z - y| \int_{\tau_a}^{\tau_e} \left| \int_0^T \int_{\partial\Omega} \frac{\partial p}{\partial \nu}(x,t) w_y(x,t;\tau)\, d\sigma(x)\, dt \right| d\tau, \tag{95}$$

which gives an approximation of $\delta^2 A$.

In two dimensions, one should rather consider the probe wave given by

$$w_\theta(x,t;\tau) = \delta\left(t + \tau - \frac{\langle x, \theta\rangle}{c}\right), \tag{96}$$

where θ is a unit vector and τ is a parameter satisfying

$$\tau > \max_{x\in\Omega}\left(\frac{\langle x,\theta\rangle}{c}\right).$$

One can still use the function

$$\tau \mapsto \int_0^T \int_{\partial\Omega} \frac{\partial p}{\partial \nu}(x,t) w_\theta(x,t;\tau)\, d\sigma(x)\, dt$$

to probe the medium as a function of τ. This quantity is nonzero on the interval $]\tau_a, \tau_e[$, where τ_a and τ_e are defined such that planes $\langle x,\theta\rangle = c\tau$ for $\tau = \tau_a$ and τ_e hit D. Changing the direction θ and intersecting stripes gives an efficient way to reconstruct the anomalies.

By exactly the same arguments as in three dimensions, one can show that

$$\delta A \approx \frac{c_s}{4}\int_{\tau_a}^{\tau_e}\left|\int_0^T\int_{\partial\Omega}\frac{\partial p}{\partial\nu}(x,t) w_\theta(x,t;\tau)\, d\sigma(x)\, dt\right| d\tau. \tag{97}$$

The above formula can be used to estimate δA.

In the case when there are m inclusions, one first computes for each l the quantity

$$\theta_{l,\text{best}} = \operatorname*{argmax}_{\theta\in[0,\pi]}\left(\min_{j\neq l}\left|\langle z_j - z_l, \theta\rangle\right|\right)$$

and then, since along the direction $\theta_{l,\text{best}}$, the inclusion D_l is well separated from all the other inclusions, one can use formula (97) to estimate its δA_l.

Reconstruction of the Absorption Coefficient

The density $A(x)$ is related to the optical absorption coefficient distribution $\mu_a(x) = \mu_a \chi(D)$ by the equation $A(x) = \mu_a(x)\Phi(x)$, where Φ is the fluence rate. The function Φ depends on the distribution of scattering and absorption within Ω, as well as the light sources. Based on the diffusion approximation to the transport equation, Φ satisfies

$$\left(\frac{i\omega}{c} + \mu_a(x) - \frac{1}{3}\nabla\cdot\frac{1}{\mu_a(x)+\mu_s(x)}\nabla\right)\Phi(x) = 0 \quad \text{in } \Omega, \tag{98}$$

with the boundary condition

$$\frac{1}{\mu_s}\frac{\partial\Phi}{\partial\nu} = g \quad \text{on } \partial\Omega. \tag{99}$$

Here, g denotes the light source, ω a given frequency, c the speed of light, and μ_s the scattering coefficient. The diffusion approximation holds when $\mu_s \gg \mu_a$.

Suppose that $d = 3$ and μ_s is known a priori. Define Φ_0 by

$$\left(\frac{i\omega}{c} - \frac{1}{3}\nabla\cdot\frac{1}{\mu_s(x)}\nabla\right)\Phi_0(x) = 0 \quad \text{in } \Omega,$$

subject to the boundary condition

$$\frac{1}{\mu_s}\frac{\partial\Phi_0}{\partial\nu} = g \quad \text{on } \partial\Omega.$$

Introduce $\hat{N}_B$ to be the Newton potential given by

$$\hat{N}_B(\xi) := \int_B \Gamma(\xi - y)\,dy,$$

where $\Gamma := -1/(4\pi|x|)$ is a fundamental solution of the Laplacian in three dimensions.

Let $\alpha := \delta^2\mu_a\Phi(z) = \delta^2 A$. As shown before, α can be reconstructed from $\frac{\partial p}{\partial \nu}$. To extract $\delta^2\mu_a$ from α, one uses the following theorem:

Theorem 10 (Fluence rate perturbations). *If B is the unit sphere, then the following expansion holds:*

$$(\Phi - \Phi_0)(z) \approx 3\alpha\mu_s(z)\hat{N}_B(0), \tag{100}$$

from which it follows that the (normalized) absorption coefficient can be approximated by

$$\delta^2\mu_a \approx \frac{\alpha}{3\alpha\mu_s(z)\hat{N}_B(0) + \Phi_0(z)}. \tag{101}$$

Separating δ from μ_a requires boundary measurements of Φ on $\partial\Omega$. One can use

$$\int_{\partial\Omega} g(\Phi - \Phi_0)\,d\sigma \approx \mu_a\Phi_0^2(z)|D| \tag{102}$$

to separately recover δ from μ_a.

In the case where μ_s is unknown, an algorithm to extract the absorption coefficient μ_a from absorbed energies obtained at multiple wavelengths was developed in [10]. It assumes that the wavelength dependence of the scattering and absorption coefficients are known. In biological tissues, the wavelength dependence of the scattering often approximates to a power law.

Bibliography and Open Questions

Basic physical principles of the photo-acoustic effect have been described, for instance, in [49, 98]. The results of this section are from [9, 10]. The location search algorithm described in this section can be extended to the case with limited-view measurements. The half-space problem has been considered [96]. In free space, one refers to [1, 2, 4, 55, 56, 70, 84] for uniqueness of the reconstruction and inversion procedures based on the spherical Radon transform. Reconstruction methods with incomplete data have been developed in [97]. Sensitivity analysis of a photo-acoustic wave to the presence of small absorbing objects has been provided in [49].

In connection with photo-acoustic imaging, it is worth mentioning the multi-physics imaging technique proposed in [52], which combines electrical impedance tomography with acoustic tomography. This method makes use of the fact that the absorbed electrical energy causes thermoelastic expansion of the tissue, which leads to propagation of a pressure wave. With the notation of Sect. 5, the induced signal is measured on the boundary of the object and can be used for calculating the absorbed electrical energy, $\mathcal{E} = \gamma|\nabla u|^2$, inside the body, from which the electrical conductivity γ can be reconstructed using, for instance, the substitution algorithm.

As for magneto-acoustic imaging with magnetic induction, it would be very interesting to design a robust inversion algorithm when the acoustic speed fluctuates randomly.

9 Conclusion

In this chapter, applications of asymptotic analysis in emerging medical imaging are outlined. This method leads to very effective and robust reconstruction algorithms in many imaging problems. Of particular interest are emerging multi-physics or hybrid-imaging approaches. These approaches allow one to overcome the severe ill-posedness character of image reconstruction. It would be very interesting to analytically investigate their robustness, with respect to incomplete data, measurement, and medium noises. Another important problem is to take into account the effect of anisotropy, dissipation, or attenuation in biological tissues.

Cross-References

- ▸ Electrical Impedance Tomography
- ▸ Optical Imaging
- ▸ Photoacoustic and Thermoacoustic Tomography: Image Formation Principles
- ▸ Mathematics of Photoacoustic and Thermoacoustic Tomography
- ▸ Tomography
- ▸ Wave Phenomena

References

1. Agranovsky, M., Kuchment, P.: Uniqueness of reconstruction and an inversion procedure for thermoacoustic and photoacoustic tomography with variable sound speed. Inverse Probl. **23**, 2089–2102 (2007)
2. Agranovsky, M., Kuchment, P., Kunyansky, L.: On reconstruction formulas and algorithms for the thermoacoustic and photoacoustic tomography. In: Wang, L.H. (ed.) Photoacoustic Imaging and Spectroscopy, pp. 89–101. CRC, Boca Raton (2009)
3. Amalu, W.C., Hobbins, W.B., Elliot, R.L.: Infrared imaging of the breast – an overview. In: Bronzino, J.D. (ed.) Medical Devices and Systems, the Biomedical Engineering Handbook, chap. 25, 3rd edn. CRC, Baton Rouge (2006)
4. Ambartsoumian, G., Patch, S.: Thermoacoustic tomography – implementation of exact back-projection formulas (2005). math.NA/0510638
5. Ammari, H.: An inverse initial boundary value problem for the wave equation in the presence of imperfections of small volume. SIAM J. Control Optim. **41**, 1194–1211 (2002)
6. Ammari, H.: An Introduction to Mathematics of Emerging Biomedical Imaging. Mathématiques and Applications, vol. 62. Springer, Berlin (2008)
7. Ammari, H., Asch, M., Guadarrama Bustos, L., Jugnon, V., Kang, H.: Transient wave imaging with limited-view data (submitted)
8. Ammari, H., Bonnetier, E., Capdeboscq, Y., Tanter, M., Fink, M.: Electrical impedance tomography by elastic deformation. SIAM J. Appl. Math. **68**, 1557–1573 (2008)
9. Ammari, H., Bossy, E., Jugnon, V., Kang, H.: Mathematical modelling in photo-acoustic imaging of small absorbers. SIAM Rev. (to appear)
10. Ammari, H., Bossy, E., Jugnon, V., Kang, H.: Quantitative photoacoustic imaging of small absorbers (submitted)
11. Ammari, H., Capdeboscq, Y., Kang, H., Kozhemyak, A.: Mathematical models and reconstruction methods in magneto-acoustic imaging. Eur. J. Appl. Math. **20**, 303–317 (2009)
12. Ammari, H., Garapon, P., Guadarrama Bustos, L., Kang, H.: Transient anomaly imaging by the acoustic radiation force. J. Differ. Equ. (to appear)
13. Ammari, H., Garapon, P., Jouve, F.: Separation of scales in elasticity imaging: a numerical study. J. Comput. Math. **28**, 354–370 (2010)
14. Ammari, H., Garapon, P., Kang, H., Lee, H.: A method of biological tissues elasticity reconstruction using magnetic resonance elastography measurements. Q. Appl. Math. **66**, 139–175 (2008)
15. Ammari, H., Garapon, P., Kang, H., Lee, H.: Effective viscosity properties of dilute suspensions of arbitrarily shaped particles (submitted)
16. Ammari, H., Griesmaier, R., Hanke, M.: Identification of small inhomogeneities: asymptotic factorization. Math. Comput. **76**, 1425–1448 (2007)
17. Ammari, H., Iakovleva, E., Kang, H., Kim, K.: Direct algorithms for thermal imaging of small inclusions. SIAM Multiscale Model. Simul. **4**, 1116–1136 (2005)
18. Ammari, H., Iakovleva, E., Lesselier, D.: Two numerical methods for recovering small electromagnetic inclusions from scattering amplitude at a fixed frequency. SIAM J. Sci. Comput. **27**, 130–158 (2005)
19. Ammari, H., Iakovleva, E., Lesselier, D.: A MUSIC algorithm for locating small inclusions buried in a half-space from the scattering amplitude at a fixed frequency. SIAM Multiscale Model. Simul. **3**, 597–628 (2005)
20. Ammari, H., Iakovleva, E., Lesselier, D., Perrusson, G.: A MUSIC-type electromagnetic imaging of a collection of small three-dimensional inclusions. SIAM J. Sci. Comput. **29**, 674–709 (2007)
21. Ammari, H., Kang, H.: High-order terms in the asymptotic expansions of the steady-state voltage potentials in the presence of conductivity inhomogeneities of small diameter. SIAM J. Math. Anal. **34**, 1152–1166 (2003)

22. Ammari, H., Kang, H.: Reconstruction of Small Inhomogeneities from Boundary Measurements. Lecture Notes in Mathematics, vol. 1846. Springer, Berlin (2004)
23. Ammari, H., Kang, H.: Boundary layer techniques for solving the Helmholtz equation in the presence of small inhomogeneities. J. Math. Anal. Appl. **296**, 190–208 (2004)
24. Ammari, H., Kang, H.: Reconstruction of elastic inclusions of small volume via dynamic measurements. Appl. Math. Optim. **54**, 223–235 (2006)
25. Ammari, H., Kang, H.: Polarization and Moment Tensors: With Applications to Inverse Problems and Effective Medium Theory. Applied Mathematical Sciences, vol. 162. Springer, New York (2007)
26. Ammari, H., Kang, H., Lee, H.: A boundary integral method for computing elastic moment tensors for ellipses and ellipsoids. J. Comput. Math. **25**, 2–12 (2007)
27. Ammari, H., Kang, H., Nakamura, G., Tanuma, K.: Complete asymptotic expansions of solutions of the system of elastostatics in the presence of an inclusion of small diameter and detection of an inclusion. J. Elast. **67**, 97–129 (2002)
28. Ammari, H., Khelifi, A.: Electromagnetic scattering by small dielectric inhomogeneities. J. Math. Pures Appl. **82**, 749–842 (2003)
29. Ammari, H., Kozhemyak, A., Volkov, D.: Asymptotic formulas for thermography based recovery of anomalies. Numer. Math. TMA **2**, 18–42 (2009)
30. Ammari, H., Kwon, O., Seo, J.K., Woo, E.J.: Anomaly detection in Tscan trans-admittance imaging system. SIAM J. Appl. Math. **65**, 252–266 (2004)
31. Ammari, H., Seo, J.K.: An accurate formula for the reconstruction of conductivity inhomogeneities. Adv. Appl. Math. **30**, 679–705 (2003)
32. Assenheimer, M., Laver-Moskovitz, O., Malonek, D., Manor, D., Nahliel, U., Nitzan, R., Saad, A.: The T-scan technology: electrical impedance as a diagnostic tool for breast cancer detection. Physiol. Meas. **22**, 1–8 (2001)
33. Bardos, C.: A mathematical and deterministic analysis of the time-reversal mirror. In: Inside Out: Inverse Problems and Applications. Mathematical Science Research Institute Publication, vol. 47, pp. 381–400. Cambridge University of Press, Cambridge (2003)
34. Bardos, C., Lebeau, G., Rauch, J.: Sharp sufficient conditions for the observation, control, and stabilization of waves from the boundary. SIAM J. Control Optim. **30**, 1024–1065 (1992)
35. Bercoff, J., Tanter, M., Fink, M.: Supersonic shear imaging: a new technique for soft tissue elasticity mapping. IEEE Trans. Ultrason. Ferroelectr. Freq. Control **51**, 396–409 (2004)
36. Bercoff, J., Tanter, M., Fink, M.: The role of viscosity in the impulse diffraction field of elastic waves induced by the acoustic radiation force. IEEE Trans. Ultrason. Ferroelectr. Freq. Control **51**, 1523–1536 (2004)
37. Borcea, L., Papanicolaou, G.C., Tsogka, C., Berrymann, J.G.: Imaging and time reversal in random media. Inverse Probl. **18**, 1247–1279 (2002)
38. Brühl, M., Hanke, M., Vogelius, M.S.: A direct impedance tomography algorithm for locating small inhomogeneities. Numer. Math. **93**, 635–654 (2003)
39. Capdeboscq, Y., De Gournay, F., Fehrenbach, J., Kavian, O.: An optimal control approach to imaging by modification. SIAM J. Imaging Sci. **2**, 1003–1030 (2009)
40. Capdeboscq, Y., Kang, H.: Improved bounds on the polarization tensor for thick domains. In: Inverse Problems, Multi-scale Analysis and Effective Medium Theory. Contemporary Mathematics, vol. 408, pp. 69–74. American Mathematical Society, Providence (2006)
41. Capdeboscq, Y., Kang, H.: Improved Hashin-Shtrikman bounds for elastic moment tensors and an application. Appl. Math. Optim. **57**, 263–288 (2008)
42. Capdeboscq, Y., Vogelius, M.S.: A general representation formula for the boundary voltage perturbations caused by internal conductivity inhomogeneities of low volume fraction. Math. Model. Numer. Anal. **37**, 159–173 (2003)
43. Capdeboscq, Y., Vogelius, M.S.: Optimal asymptotic estimates for the volume of internal inhomogeneities in terms of multiple boundary measurements. Math. Model. Numer. Anal. **37**, 227–240 (2003)
44. Cedio-Fengya, D.J., Moskow, S., Vogelius, M.S.: Identification of conductivity imperfections of small diameter by boundary measurements: continuous dependence and computational reconstruction. Inverse Probl. **14**, 553–595 (1998)

45. Chambers, D.H., Berryman, J.G.: Analysis of the time-reversal operator for a small spherical scatterer in an electromagnetic field. IEEE Trans. Antennas Propag. **52**, 1729–1738 (2004)
46. Chambers, D.H., Berryman, J.G.: Time-reversal analysis for scatterer characterization. Phys. Rev. Lett. **92**, 023902–1 (2004)
47. Devaney, A.J.: Time reversal imaging of obscured targets from multistatic data. IEEE Trans. Antennas Propag. **523**, 1600–1610 (2005)
48. Fink, M.: Time-reversal acoustics. Contemp. Math. **408**, 151–179 (2006)
49. Fisher, A.R., Schissler, A.J., Schotland, J.C.: Photoacoustic effect of multiply scattered light. Phys. Rev. E **76**, 036604 (2007)
50. Fouque, J.P., Garnier, J., Papanicolaou, G., Solna, K.: Wave Propagation and Time Reversal in Randomly Layered Media. Springer, New York (2007)
51. Friedman, A., Vogelius, M.S.: Identification of small inhomogeneities of extreme conductivity by boundary measurements: a theorem on continuous dependence. Arch. Ration. Mech. Anal. **105**, 299–326 (1989)
52. Gebauer, B., Scherzer, O.: Impedance-acoustic tomography. SIAM J. Appl. Math. **69**, 565–576 (2008)
53. Greenleaf, J.F., Fatemi, M., Insana, M.: Selected methods for imaging elastic properties of biological tissues. Annu. Rev. Biomed. Eng. **5**, 57–78 (2003)
54. Haider, S., Hrbek, A., Xu, Y.: Magneto-acousto-electrical tomography: a potential method for imaging current density and electrical impedance. Physiol. Meas. **29**, 41–50 (2008)
55. Haltmeier, M., Scherzer, O., Burgholzer, P., Nuster, R., Paltauf, G.: Thermoacoustic tomography and the circular Radon transform: exact inversion formula. Math. Model. Methods Appl. Sci. **17**(4), 635–655 (2007)
56. Haltmeier, M., Schuster, T., Scherzer, O.: Filtered backprojection for thermoacoustic computed tomography in spherical geometry. Math. Methods Appl. Sci. **28**, 1919–1937 (2005)
57. Hanke, M.: On real-time algorithms for the location search of discontinuous conductivities with one measurement. Inverse Probl. **24**, 045005 (2008)
58. Harrach, B., Seo, J.K.: Detecting inclusions in electrical impedance tomography without reference measurements. SIAM J. Appl. Math. **69**, 1662–1681 (2009)
59. Isakov, V.: Inverse Problems for Partial Differential Equations. Applied Mathematical Sciences, vol. 127. Springer, New York (1998)
60. Kang, H., Kim, E., Kim, K.: Anisotropic polarization tensors and determination of an anisotropic inclusion. SIAM J. Appl. Math. **65**, 1276–1291 (2003)
61. Kang, H., Seo, J.K.: Layer potential technique for the inverse conductivity problem. Inverse Probl. **12**, 267–278 (1996)
62. Kang, H., Seo, J.K.: Identification of domains with near-extreme conductivity: global stability and error estimates. Inverse Probl. **15**, 851–867 (1999)
63. Kang, H., Seo, J.K.: Inverse conductivity problem with one measurement: uniqueness of balls in R3. SIAM J. Appl. Math. **59**, 1533–1539 (1999)
64. Kang, H., Seo, J.K.: Recent progress in the inverse conductivity problem with single measurement. In: Inverse Problems and Related Fields, pp. 69–80. CRC, Boca Raton (2000)
65. Kim, Y.J., Kwon, O., Seo, J.K., Woo, E.J.: Uniqueness and convergence of conductivity image reconstruction in magnetic resonance electrical impedance tomography. Inverse Probl. **19**, 1213–1225 (2003)
66. Kim, S., Kwon, O., Seo, J.K., Yoon, J.R.: On a nonlinear partial differential equation arising in magnetic resonance electrical impedance imaging. SIAM J. Math. Anal. **34**, 511–526 (2002)
67. Kim, S., Lee, J., Seo, J.K., Woo, E.J., Zribi, H.: Multifrequency transadmittance scanner: mathematical framework and feasibility. SIAM J. Appl. Math. **69**, 22–36 (2008)
68. Kohn, R., Vogelius, M.: Identification of an unknown conductivity by means of measurements at the boundary. In: McLaughlin, D. (ed.) Inverse Problems. SIAM-AMS Proceedings, vol. 14, pp. 113–123. American Mathmetical Society, Providence (1984)
69. Kolehmainen, V., Lassas, M., Ola, P.: The inverse conductivity problem with an imperfectly known boundary. SIAM J. Appl. Math. **66**, 365–383 (2005)
70. Kuchment, P., Kunyansky, L.: Mathematics of thermoacoustic tomography. Euro. J. Appl. Math. **19**, 191–224 (2008)

71. Kuchment, P., Kunyansky, L.: Synthetic focusing in ultrasound modulated tomography. Inverse Probl. Imaging (to appear)
72. Kwon, O., Seo, J.K.: Total size estimation and identification of multiple anomalies in the inverse electrical impedance tomography. Inverse Probl. **17**, 59–75 (2001)
73. Kwon, O., Seo, J.K., Yoon, J.R.: A real-time algorithm for the location search of discontinuous conductivities with one measurement. Commun. Pure Appl. Math. **55**, 1–29 (2002)
74. Kwon, O., Yoon, J.R., Seo, J.K., Woo, E.J., Cho, Y.G.: Estimation of anomaly location and size using impedance tomography. IEEE Trans. Biomed. Eng. **50**, 89–96 (2003)
75. Li, X., Xu, Y., He, B.: Magnetoacoustic tomography with magnetic induction for imaging electrical impedance of biological tissue. J. Appl. Phys. **99**, Art. No. 066112 (2006)
76. Li, X., Xu, Y., He, B.: Imaging electrical impedance from acoustic measurements by means of magnetoacoustic tomography with magnetic induction (MAT-MI). IEEE Trans. Biomed. Eng. **54**, 323–330 (2007)
77. Lipton, R.: Inequalities for electric and elastic polarization tensors with applications to random composites. J. Mech. Phys. Solids **41**, 809–833 (1993)
78. Manduca, A., Oliphant, T.E., Dresner, M.A., Mahowald, J.L., Kruse, S.A., Amromin, E., Felmlee, J.P., Greenleaf, J.F., Ehman, R.L.: Magnetic resonance elastography: non-invasive mapping of tissue elasticity. Med. Image Anal. **5**, 237–254 (2001)
79. Milton, G.W.: The Theory of Composites. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, Cambridge (2001)
80. Montalibet, A., Jossinet, J., Matias, A., Cathignol, D.: Electric current generated by ultrasonically induced Lorentz force in biological media. Med. Biol. Eng. Comput. **39**, 15–20 (2001)
81. Muthupillai, R., Lomas, D.J., Rossman, P.J., Greenleaf, J.F., Manduca, A., Ehman, R.L.: Magnetic resonance elastography by direct visualization of propagating acoustic strain waves. Science **269**, 1854–1857 (1995)
82. Mast, T.D., Nachman, A., Waag, R.C.: Focusing and imagining using eigenfunctions of the scattering operator. J. Acoust. Soc. Am. **102**, 715–725 (1997)
83. Parisky, Y.R., Sardi, A., Hamm, R., Hughes, K., Esserman, L., Rust, S., Callahan, K.: Efficacy of computerized infrared imaging analysis to evaluate mammographically suspicious lesions. Am. J. Radiol. **180**, 263–269 (2003)
84. Patch, S.K., Scherzer, O.: Guest editors' introduction: photo- and thermo-acoustic imaging. Inverse Probl. **23**, S1–S10 (2007)
85. Pernot, M., Montaldo, G., Tanter, M., Fink, M.: "Ultrasonic stars" for time-reversal focusing using induced cavitation bubbles. Appl. Phys. Lett. **88**, 034102 (2006)
86. Pinker, S.: How the Mind Works. Penguin Science, Harmondsworth (1997)
87. Prada, C., Thomas, J.-L., Fink, M.: The iterative time reversal process: analysis of the convergence. J. Acoust. Soc. Am. **97**, 62–71 (1995)
88. Seo, J.K., Kwon, O., Ammari, H., Woo, E.J.: Mathematical framework and anomaly estimation algorithm for breast cancer detection using TS2000 configuration. IEEE Trans. Biomed. Eng. **51**, 1898–1906 (2004)
89. Seo, J.K., Woo, E.J.: Multi-frequency electrical impedance tomography and magnetic resonance electrical impedance tomography. In: Mathematical Modeling in Biomedical Imaging I. Lecture Notes in Mathematics: Mathematical Biosciences Subseries, vol. 1983. Springer, Berlin (2009)
90. Sinkus, R., Siegmann, K., Xydeas, T., Tanter, M., Claussen, C., Fink, M.: MR elastography of breast lesions: understanding the solid/liquid duality can improve the specificity of contrast-enhanced MR mammography. Magn. Reson. Med. **58**, 1135–1144 (2007)
91. Sinkus, R., Tanter, M., Catheline, S., Lorenzen, J., Kuhl, C., Sondermann, E., Fink, M.: Imaging anisotropic and viscous properties of breast tissue by magnetic resonance-elastography. Magn. Reson. Med. **53**, 372–387 (2005)
92. Sinkus, R., Tanter, M., Xydeas, T., Catheline, S., Bercoff, J., Fink, M.: Viscoelastic shear properties of in vivo breast lesions measured by MR elastography. Magn. Reson. Imaging **23**, 159–165 (2005)

93. Tanter, M., Fink, M.: Time reversing waves for biomedical applications. In: Mathematical Modeling in Biomedical Imaging I. Lecture Notes in Mathematics: Mathematical Biosciences Subseries, vol. 1983. Springer, Berlin (2009)
94. Therrien, C.W.: Discrete Random Signals and Statistical Signal Processing. Prentice-Hall, Englewood Cliffs (1992)
95. Vogelius, M.S., Volkov, D.: Asymptotic formulas for perturbations in the electromagnetic fields due to the presence of inhomogeneities. Math. Model. Numer. Anal. **34**, 723–748 (2000)
96. Wang, L.V., Yang, X.: Boundary conditions in photoacoustic tomography and image reconstruction. J. Biomed. Opt. **12**, 014027 (2007)
97. Xu, Y., Wang, L.V.: Ambartsoumian, G., Kuchment, P.: Reconstructions in limited view thermoacoustic tomography. Med. Phys. **31**, 724–733 (2004)
98. Xu, M., Wang, L.V.: Photoacoustic imaging in biomedicine. Rev. Sci. Instrum. **77**, 041101 (2006)

Sampling Methods

Martin Hanke-Bourgeois and Andreas Kirsch

Contents

1 Introduction 592
2 The Factorization Method in Impedance Tomography 594
Impedance Tomography in the Presence of Insulating Inclusions 595
Conducting Obstacles 603
Local Data 611
Other Generalizations 612
3 The Factorization Method in Inverse Scattering Theory 615
Inverse Acoustic Scattering by a Sound-Soft Obstacle 616
Inverse Electromagnetic Scattering by an Inhomogeneous Medium 622
Historical Remarks and Open Questions 627
4 Related Sampling Methods 628
The Linear Sampling Method 628
MUSIC 631
The Singular Sources Method 635
The Probe Method 637
5 Conclusion 640
6 Appendix 640
Cross-References 643
References 643

M. Hanke-Bourgeois (✉)
Institut für Mathematik, Johannes Gutenberg-Universität Mainz, Mainz, Germany
e-mail: hanke@mathematik.uni-mainz.de

A. Kirsch
Department of Mathematics, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany
e-mail: andreas.kirsch@kit.edu

© Springer Science+Business Media New York 2015

O. Scherzer (ed.), *Handbook of Mathematical Methods in Imaging*,
DOI 10.1007/978-1-4939-0790-8_12

Abstract

The topic of this chapter is devoted to *shape identification problems*, i.e., problems where the shape of an object has to be determined from indirect measurements. In contrast to iterative methods where a sequence of forward problems has to be computed the *sampling methods* avoid the (usually expansive) computation of the forward problems. Instead, a class of test objects (e.g., points) are chosen and a binary criterium is constructed which depends on the measured data only, and which decides whether this test object is inside or outside of the searched for domain. In this chapter, the *factorization method* is explained for the impedance tomography problem with insulating or conducting inclusions, for scattering theory for time harmonic acoustic plane waves in the presence of a perfectly sound–soft obstacle, and for electromagnetic scattering by an inhomogeneous conducting medium. Brief descriptions of related sampling methods, such as the *linear sampling method*, *MUSIC*, the *singular sources method*, and the *probe method* complement this chapter.

1 Introduction

The topic of this chapter is devoted to *shape identification problems*, i.e., problems where the shape of an object has to be determined from indirect measurements. Such a situation typically occurs in problems of *tomography*, in particular electrical impedance tomography or optical tomography. For example, a current through a homogeneous object will in general induce a different potential than the same current through the same object containing an enclosed cavity. In impedance tomography, the task is to determine the shape of the cavity from measurements of the potential on the boundary of the object. For survey articles on this subject, we refer to [18, 54], and chapter ▸ Electrical Impedance Tomography in this volume.

As a second of these fields, we mention *inverse scattering problems* where one wants to detect – and identify – unknown objects through the use of acoustic, electromagnetic, or elastic waves. Similar to the above, one of the important problems in inverse scattering theory is to determine the shape of the scattering obstacle from field measurements. Applications of inverse scattering problems occur in such diverse areas as medical imaging, material science, nondestructive testing, radar, remote sensing, or seismic exploration. A survey on the state of the art of the mathematical theory and numerical approaches for solving inverse time-harmonic scattering problems until 1998 can be found in the standard monograph [36] (see also chapter ▸ Inverse Scattering or [83] for an introduction and survey on inverse scattering problems).

Shape identification problems are intrinsically *nonlinear*, i.e., the measured quantities do not depend linearly on the shape. Even the notion of linearity does not make sense since, in general, the set of admissible shapes does not carry a linear structure. Traditional (and still very successful) approaches describe the objects by appropriate parameterizations and compute the parameters by *iterative schemes* as, e.g., Newton-type methods. Newton-type methods are attractive because of their

fast convergence, although they require a good initial guess to converge. Still, these methods are widely used – partly because techniques from shape optimization theory can be used to characterize the required first- or second-order derivatives. We refer to [84, 89] for general references and to [57, 58, 65] for applications in inverse scattering theory.

While classical iterative algorithms use explicit parameterizations of the object, new shape optimization methods have been developed since around 1995 which completely avoid the use of parameterizations and replace the classical Fréchet derivative by a geometrically motivated *topological derivative*, see, e.g., [50] for the application of these methods in the inverse scattering context. Yet these methods have the shortcoming that they are not able to change the number of connectivity components during the algorithm. This has led to the development of *level set methods* which are based on implicit representations of the unknown object involving an "evolution parameter" t. We refer to [25] or chapter ▸ Level Set Methods for Structural Inversion and Image Reconstruction for recent surveys.

While very successful in many cases, iterative methods for shape identification problems – may they use classical tools as the Fréchet derivative or more recent techniques such as domain derivatives, level curves, or topological derivatives – are computationally very expensive since they require the solution of a direct problem in every step. Furthermore, for many important cases, the convergence theory is still missing. This is due to the fact that these problems are not only nonlinear but also because their linearizations are *improperly posed*. Although there exist many results on the convergence of (regularized) iterative methods for solving nonlinear improperly posed problems (see, e.g., [39, 64] or chapter ▸ Iterative Solution Methods), the assumptions for convergence are not met in the applications to shape identification problems. (Or, at least, it is unknown whether these assumptions are fulfilled or not.)

These difficulties and disadvantages of iterative schemes gave rise to the development of different classes of *non-iterative* methods which avoid the solution of a sequence of direct problems. We briefly mention *decomposition methods* (according to the notion of [37]) which consist of an analytic continuation step (which is linear but highly improperly posed) and a nonlinear step of finding the boundary of the unknown domain by forcing the boundary condition to hold. We refer to section "Decomposition Methods" in chapter ▸ Inverse Scattering.

This chapter will focus on a different class of non-iterative methods, the so-called *sampling methods*. The common idea of these methods is the construction of criteria on the known data to decide whether a given test object (a point or a curve or a set) is inside or outside the unknown domain D. Then, a grid of "sampling" points is chosen in a region that is known to contain the unknown domain D, in order to compute the (approximate) characteristic function of D. The different kinds of sampling methods differ in the way of defining the criteria and in the type of test objects.

One of the first methods which falls into this class has been developed by David Colton and one of the authors (A. K.) in 1996 [35], now known as the *linear sampling method*. Its origin goes back to the *dual space method* developed

between 1985 and 1990 (see, e.g., [36]). The numerical implementation of the linear sampling method is extremely simple and fast because sampling is done by points z only. For every sampling point z, one has to compute the field of a point source in z with respect to the background medium (essentially, one has to compute the fundamental solution of the underlying differential operator; if the background is constant, the response is given analytically) and evaluate a series, i.e., a finite sum in practice.

A problem with the linear sampling method from the mathematical point of view is that the computable criterion is only a sufficient condition which is, in general, not necessary. The *factorization method* overcomes this drawback and yields a criterion for z which is both necessary and sufficient. Therefore, this method succeeds to provide a simple formula for the characteristic function of D which can easily be used for numerical computations.

The factorization method consists of three components. First, a "measurement operator" M is factorized in three factors of the form

$$M = AGA^*, \tag{1}$$

where A^* is the dual operator of A with respect to the L^2 topology. Second, the range of A is characterized by the obstacle D, and vice versa. Third, if the operator G satisfies a certain coercivity condition, then the range of A can be determined by the given operator M. This requires some functional analytic results on range identities which we have collected in an appendix.

Combining these three steps yields an explicit characterization of the unknown obstacle D by the measurement operator M.

The outline of this chapter is as follows. First, in Sect. 2, we present the factorization method for two different settings in the impedance tomography context. In the very first setting, we deal with insulating inclusions, and this allows for a very elementary presentation of the method. Afterwards, in Sect. 3, we turn to applications from inverse acoustic and (full 3D) electromagnetic scattering. Finally, we give a brief overview of other sampling type methods in Sect. 4, including the original linear sampling method and MUSIC-type methods.

2 The Factorization Method in Impedance Tomography

We start with the impedance tomography problem. Consider an object that fills a simply connected domain $\Omega \subset \mathbb{R}^n$ with Lipschitz continuous boundary, where $n = 2$ or $n = 3$, respectively. We assume that the object is a homogeneous and isotropic conductor, except for a finite number m of so-called inclusions, given by domains $D_i \subset \Omega$, $i = 1, \ldots, m$, with Lipschitz continuous boundaries ∂D_i. We assume that these domains are well separated, i.e., $\overline{D_i} \cap \overline{D_j} = \emptyset$ when $i \neq j$, and that the complement of the closure $\overline{D}$ of $D = \cup_{i=1}^m D_i$ is connected. In impedance tomography, currents are imposed through the boundary of the object,

and the resulting boundary potentials are measured. Linear independent boundary currents yield independent pieces of information, which can be used as input data to determine the unknown shapes and positions of the inclusions.

In practice, at least in most medical applications, the boundary currents have a frequency in the kHz range (5–500 kHz), and the dc approximation with a positive real conductivity σ (or possibly a positive definite tensor) serves as a suitable physical model. Without loss of generality, we can always assume that the homogeneous conductivity of the object equals $\sigma = 1$, whereas $\sigma \neq 1$ within the inclusions.

Below we will consider two specific scenarios. In the first one, we assume that the inclusions are insulating, formally corresponding to the case where $\sigma = 0$. Our analysis of the factorization method for the corresponding inverse problem will be somewhat nonstandard; in particular, we employ a factorization in only two factors instead of three as in (1), but this allows for a most elementary treatment of the method.

Subsequently, we show how to deal with conducting obstacles with a conductivity tensor σ. Of particular interest is the setting where the object under consideration can be modeled as a half space; examples of this sort arise in geophysics, cf. [78], and in medicine, e.g., when a planar device is used for mammography examinations, cf. [92]. Another interesting application for the half-space problem has recently been considered in [17]. We therefore briefly describe the differences that arise in this context (mainly in the theoretical justification of the method).

We conclude our case studies with a setting where the inclusion degenerates to a crack, i.e., an $n-1$ dimensional smooth manifold within Ω. This application requires some care in the appropriate implementation of the factorization method.

Impedance Tomography in the Presence of Insulating Inclusions

To begin with, we take up the case where Ω is a bounded domain, and the domains $D_i \subset \Omega$, $i = 1, \ldots, m$ correspond to insulating inclusions. Within the dc model, the potential u_0 induced by a boundary current f is given by

$$\begin{aligned} \Delta u_0 = 0 \quad \text{in } \Omega \backslash \overline{D}, \qquad \frac{\partial}{\partial \nu} u_0 = 0 \quad \text{on } \partial D, \\ \frac{\partial}{\partial \nu} u_0 = f \quad \text{on } \partial\Omega, \qquad \int_{\partial\Omega} u_0 \, \mathrm{d}s = 0, \end{aligned} \tag{2}$$

where the normal vectors ν on $\partial\Omega$ and ∂D are pointing into the exterior of Ω and D, respectively. In order to make the forward problem (2) well posed, we restrict f to be square integrable with vanishing mean on $\partial\Omega$. The corresponding set of admissible boundary currents is

$$L^2_\diamond(\partial\Omega) = \left\{ f \in L^2(\partial\Omega) : \int_{\partial\Omega} f \, \mathrm{d}s = 0 \right\}. \tag{3}$$

Under these assumptions, problem (2) has a unique (weak) solution

$$u_0 \in H^1_\diamond(\Omega\backslash\overline{D}) = \left\{ u \in H^1(\Omega\backslash\overline{D}) : \int_{\partial\Omega} u \, \mathrm{d}s = 0 \right\}.$$

The last condition in (2) normalizes this boundary potential to have vanishing mean; without this condition, the solution would only be unique up to additive constants, reflecting the fact that only the voltage, i.e., the difference between the potential at two different points, is a well-defined physical quantity.

Therefore, the **direct problem** is to determine the field u_0 when f and D are given.

The quantity that is measured in impedance tomography is the trace $g_0 = u_0|_{\partial\Omega}$, i.e., the boundary potential. The corresponding measurement operator

$$\Lambda_0 : \begin{cases} L^2_\diamond(\partial\Omega) \to L^2_\diamond(\partial\Omega), \\ f \mapsto g_0 = u_0|_{\partial\Omega}, \end{cases} \tag{4}$$

i.e., the so-called *Neumann-Dirichlet operator*, is usually referred to as *absolute data* in impedance tomography.

The **inverse problem** is to determine the shape of D from the measurement operator Λ_0.

For the factorization method, we employ *relative data*, that is, the difference between the above Neumann-Dirichlet operator and the corresponding one for a completely homogeneous object in Ω. To be precise, let $u_\mathbb{1}$ be the reference solution for the homogeneous object, given the same boundary current $f \in L^2_\diamond(\partial\Omega)$,

$$\Delta u_\mathbb{1} = 0 \quad \text{in } \Omega, \qquad \frac{\partial}{\partial \nu} u_\mathbb{1} = f \quad \text{on } \partial\Omega, \qquad \int_{\partial\Omega} u_\mathbb{1} \, \mathrm{d}s = 0, \tag{5}$$

and denote by $\Lambda_\mathbb{1} : f \mapsto g_\mathbb{1} = u_\mathbb{1}|_{\partial\Omega}$ the Neumann-Dirichlet map associated with (5). It is the relative data $M = \Lambda_0 - \Lambda_\mathbb{1}$ that later enters in (1) to lay the grounds for the setting of the factorization method.

We refer to chapter ▸ Electrical Impedance Tomography for a more elaborate treatment of the impedance tomography problem, but we will see below that $\Lambda_0 - \Lambda_\mathbb{1}$ is a bounded and positive self-adjoint operator. We also do not discuss practical issues such as electrode models that should be incorporated into a realistic problem setting. For the same reason, we do not comment on how to obtain relative data in practice; the generation of accurate reference data is indeed a difficult subject, and some work-arounds have therefore been suggested for this purpose. (We like to highlight one recent approach from [56], where different frequencies are used in the experimental setup to obtain relative data. This approach, however, leads to a different variant of the factorization method than the one that is described here.) Our specification of the impedance tomography problem is thus a purely mathematical one, although it can be shown to be a pretty reasonable approximation of the real case, cf., e.g., [60, 77].

Before we continue, we pause to comment on the nature of the relative data introduced above. Any function h in the range $\mathcal{R}(\Lambda_0 - \Lambda_{\mathbb{1}})$ of $\Lambda_0 - \Lambda_{\mathbb{1}}$ corresponds to a suitable input current $f \in L^2_\diamond(\partial\Omega)$, such that h is the trace of $w = u_0 - u_{\mathbb{1}} : \Omega\backslash\overline{D} \to \mathbb{R}$, where u_0 and $u_{\mathbb{1}}$ are the solutions of (2) and (5), respectively. As u_0 and $u_{\mathbb{1}}$ are both harmonic in $\Omega\backslash\overline{D}$, the same holds true for w; on top of that, like u_0 and $u_{\mathbb{1}}$, w has finite H^1 norm on $\Omega\backslash\overline{D}$, as well as vanishing mean on $\partial\Omega$. Moreover, w has homogeneous Neumann boundary conditions on $\partial\Omega$, as u_0 and $u_{\mathbb{1}}$ both satisfy the same Neumann boundary condition. And finally, on ∂D_i, $i = 1, \ldots, m$, we have

$$\int_{\partial D_i} \frac{\partial}{\partial \nu} w \, \mathrm{d}s = -\int_{\partial D_i} \frac{\partial}{\partial \nu} u_{\mathbb{1}} \, \mathrm{d}s = 0$$

by virtue of Green's formula. Accordingly, the range of $\Lambda_0 - \Lambda_{\mathbb{1}}$ consists of traces of potentials w from

$$\mathcal{W} = \left\{ w \in H^1_\diamond(\Omega\backslash\overline{D}) : \Delta w = 0, \frac{\partial}{\partial \nu} w = 0 \text{ on } \partial\Omega, \right.$$
$$\left. \int_{\partial D_i} \frac{\partial}{\partial \nu} w \, \mathrm{d}s = 0, i = 1, \ldots, m \right\}. \tag{6}$$

It is well known that harmonic functions have infinite smoothness. Moreover, as the elements of $\mathcal{W}$ have a vanishing Neumann derivative on $\partial\Omega$, the "variation" of w on $\partial\Omega$ can only be caused by their behavior near the boundary of D – unless the boundary of Ω is non-smooth. In other words, the (local) variation of the trace of some function $w \in \mathcal{W}$ is an indicator for the (local) width of the domain $\Omega\backslash\overline{D}$. In fact, as we will show next, it is possible to characterize D completely, if the set of all traces of $\mathcal{W}$ on $\partial\Omega$ were known. (For one insulating inclusion, it is even known that the trace of one single potential $w \in \mathcal{W}$ is enough to identify D, cf., e.g., [16]. For conducting obstacles, with known conductivity, the corresponding uniqueness problem is still open.)

To this end, we introduce the Neumann function $N(\cdot, z)$ associated with the Laplacian in the domain Ω, which is given as the (distributional) solution of the problem

$$-\Delta N(x,z) = \delta(x-z) \quad \text{in } \Omega, \qquad \frac{\partial}{\partial \nu} N(x,z) = -\frac{1}{|\partial\Omega|} \quad \text{on } \partial\Omega,$$
$$\int_{\partial\Omega} N(x,z) \, \mathrm{d}s(x) = 0, \tag{7}$$

where $z \in \Omega$ is kept fixed, and the differential operators act on the x-variable only. To achieve a unique solution, we have normalized $N(\cdot, z)$ to have vanishing mean on $\partial\Omega$. The directional derivative

$$U_z(x) = p \cdot \mathrm{grad}_z \, N(x,z) \tag{8}$$

with respect to z of N in direction p (of unit length) yields the potential of a dipole source in z with moment p in the presence of an insulated boundary $\partial\Omega$: We refer to U_z as the dipole potential, tacitly assuming the dipole moment to be fixed. (All subsequent results hold true for an arbitrary choice of $p \in \mathbb{R}^n$ with $|p| = 1$, and it appears that there is still space to improve the numerical performance of the method, especially in three-space dimensions, provided that this property is exploited in an optimal way.) We remark that U_z behaves like

$$U_z(x) \sim \begin{cases} \dfrac{1}{2\pi} \dfrac{(x-z)\cdot p}{|x-z|^2}, & n = 2, \\ \dfrac{1}{4\pi} \dfrac{(x-z)\cdot p}{|x-z|^3}, & n = 3, \end{cases} \qquad \text{as } x \to z, \tag{9}$$

and, in fact, U_z agrees with the right-hand side of (9) up to a harmonic function. This statement holds true for every fixed $z \in \Omega$.

Now we are ready to formulate the characterization of the inclusion D as it has been established by Brühl in his dissertation [21] (see also [22]) and which constitutes the basis for the factorization method.

Theorem 1. *A point $z \in \Omega$ belongs to D, if and only if the trace $\phi_z = U_z|_{\partial\Omega}$ coincides with the trace of some potential $w \in \mathcal{W}$.*

Proof. First, let $z \in D$. Then the dipole potential U_z is harmonic in $\Omega\backslash\{z\}$, i.e., in $\Omega\backslash\overline{D}$ and in a neighborhood of ∂D. Accordingly U_z belongs to $H^1_\diamond(\Omega\backslash\overline{D})$. As $N(x, z)$ has the same Neumann boundary data for any $z \in \mathbb{R}^n$, its directional derivative with respect to z has vanishing Neumann data on $\partial\Omega$. Moreover, according to Green's formula,

$$\int_{\partial D_i} \frac{\partial}{\partial \nu} U_z \, \mathrm{d}s = 0 \tag{10}$$

for every component D_i of D which does not contain z; however, as the total flux of U_z across $\partial(\Omega\backslash\overline{D})$ vanishes as well, (10) must also hold true for that component D_i of D which does contain z. Therefore, $U_z \in \mathcal{W}$, and its trace belongs to the corresponding trace space.

Now, let $z \notin \overline{D}$, and assume that the trace ϕ_z of the dipole potential U_z is the trace of a potential $w \in \mathcal{W}$. As we have seen in the first part of this proof, U_z and w thus have the same Cauchy data on $\partial\Omega$, and it follows from the uniqueness of solutions of the Cauchy problem for the Poisson equation that U_z and w coincide in $\Omega\backslash(\overline{D} \cup \{z\})$, where both are harmonic. (It is here where the assumption on the connectedness of $\Omega\backslash\overline{D}$ is needed.) Now, w extends as a harmonic function into the point z and, hence, is bounded near z, whereas U_z is not, cf. (9). This provides the desired contradiction.

In the last case, where z sits on the boundary of D, we can use the same argument as before to show that w and U_z coincide in $\Omega\backslash\overline{D}$. According to (6), U_z must

therefore have a finite H^1-norm on $\Omega\backslash\overline{D}$, which contradicts the asymptotic behavior (9) near $z \in \partial D$. (This argument requires the Lipschitz continuity of ∂D, because this assumption makes sure that we can find an open-cone $\mathcal{C} \subset \Omega\backslash\overline{D}$ with vertex in z and, hence, that the integral $\int_{\mathcal{C}} |\text{grad } U_z|^2 \text{d}x$ is unbounded.) ■

It turns out that the potentials $w = u_0 - u_{\mathbb{1}}$, which provide the given relative data, have additional features that are not captured by the description of the set $\mathcal{W}$ of (6). For example, if the boundaries of the domains D_i are smooth, then the potential u_0 of (2) can be extended by reflection to a certain subset of D, showing that w has a harmonic extension to a larger domain than just $\Omega\backslash\overline{D}$ (see [54]; Appendix). Therefore, the space spanned by the relative data is *smaller* than the trace space of $\mathcal{W}$ in general. Still, there is a means to deduce this trace space from the given relative data – and the appropriate tool is the factorization method.

At this point we deviate from the usual presentation of the factorization method to opt for a more elementary derivation of the main results. Instead of the usual factorization of the data map in three factors as in (1), we follow the approach in [23] and factor the relative data in only two parts, namely,

$$\Lambda_0 - \Lambda_{\mathbb{1}} = K^* K, \tag{11}$$

where K^* is an appropriate adjoint of the operator K given by

$$K : f \mapsto \begin{cases} u_0 - u_{\mathbb{1}} & \text{in } \Omega\backslash\overline{D}, \\ c_i - u_{\mathbb{1}} & \text{in } \overline{D}_i, i = 1, \ldots, m, \end{cases} \tag{12}$$

and the real numbers c_i in (12) are the means of the potential u_0 at the boundaries of the insulating inclusions, i.e.,

$$c_i = \frac{1}{|\partial D_i|} \int_{\partial D_i} u_0 \, \text{d}s, \qquad i = 1, \ldots, m. \tag{13}$$

We claim (see Theorem 2 below for a proof) that K is a continuous operator from $L^2_\diamond(\partial\Omega)$ to $\mathcal{X}$, where

$$\mathcal{X} = \Big\{ v : \Omega \to \mathbb{R} \ : \ v\big|_{\Omega\backslash\overline{D}} \in H^1_\diamond(\Omega\backslash\overline{D}), v\big|_D \in H^1(D), \\ \int_{\partial D_i} [v] \, \text{d}s = 0, i = 1, \ldots, m \Big\}. \tag{14}$$

In this definition, again, the subscript $\diamond$ indicates that any $v \in \mathcal{X}$ is required to have vanishing mean on $\partial\Omega$, and

$$[v] = v^+|_{\partial D} - v^-|_{\partial D}$$

denotes the jump of v across the boundary of the inclusion(s), defined in the appropriate trace spaces. Here and below we denote by v^+ and v^- the restriction

of a generic element $v \in \mathcal{X}$ to $\Omega\backslash\overline{D}$ and D, respectively. We equip $\mathcal{X}$ with the inner product

$$\begin{aligned}(v,w)_{\mathcal{X}} &= \int_{\Omega\backslash\partial D} \operatorname{grad} v \cdot \operatorname{grad} w \mathrm{d}x \\ &= \int_D \operatorname{grad} v^- \cdot \operatorname{grad} w^- \mathrm{d}x + \int_{\Omega\backslash\overline{D}} \operatorname{grad} v^+ \cdot \operatorname{grad} w^+ \mathrm{d}x, \end{aligned} \tag{15}$$

which turns $\mathcal{X}$ into a Hilbert space. Take note that $H^1_\diamond(\Omega)$, i.e., the set of all functions from $H^1(\Omega)$ with vanishing mean on $\partial\Omega$, is a subset of $\mathcal{X}$.

Lemma 1. *Let $\mathcal{K} \subset \mathcal{X}$ be the set of all elements $w \in \mathcal{X}$ that are harmonic in $\Omega\backslash\partial D$ and satisfy*

$$\frac{\partial}{\partial \nu} w = 0 \text{ on } \partial\Omega \quad \text{and} \quad \left[\frac{\partial}{\partial \nu} w\right] = 0 \text{ on } \partial D.$$

Then $\mathcal{K}$ is the orthogonal complement of $H^1_\diamond(\Omega)$ in $\mathcal{X}$.

Proof. Using Green's formula for any $v \in H^1_\diamond(\Omega)$ and any $w \in \mathcal{X}$ that is harmonic in $\Omega\backslash\partial D$, we obtain

$$\begin{aligned}\int_{\Omega\backslash\partial D} \operatorname{grad} v \cdot \operatorname{grad} w \mathrm{d}x &= \int_{\partial\Omega} v \frac{\partial w}{\partial \nu} \mathrm{d}s - \int_{\partial D} v \frac{\partial w^+}{\partial \nu} \mathrm{d}s + \int_{\partial D} v \frac{\partial w^-}{\partial \nu} \mathrm{d}s \\ &= \int_{\partial\Omega} v \frac{\partial w}{\partial \nu} \mathrm{d}s - \int_{\partial D} v \left[\frac{\partial w}{\partial \nu}\right] \mathrm{d}s, \end{aligned} \tag{16}$$

as v has a well-defined unique trace on ∂D. Now, if we choose $w \in \mathcal{K}$, then both integrals vanish, and hence $w \perp v$ with respect to the scalar product in $\mathcal{X}$.

Vice versa, pick $w \in \mathcal{X}$ from the orthogonal complement of $H^1_\diamond(\Omega)$, and let v be a C^∞ function with compact support in $\Omega\backslash\overline{D}$, then Green's formula yields

$$\begin{aligned}\int_{\Omega\backslash\overline{D}} w \Delta v \mathrm{d}x &= \int_{\partial\Omega} w \frac{\partial v}{\partial \nu} \mathrm{d}s - \int_{\partial D} w \frac{\partial v}{\partial \nu} \mathrm{d}s - \int_{\Omega\backslash\overline{D}} \operatorname{grad} w \cdot \operatorname{grad} v \mathrm{d}x \\ &= \int_{\partial\Omega} w \frac{\partial v}{\partial \nu} \mathrm{d}s - \int_{\partial D} w \frac{\partial v}{\partial \nu} \mathrm{d}s - \int_{\Omega\backslash\partial D} \operatorname{grad} w \cdot \operatorname{grad} v \mathrm{d}x, \end{aligned}$$

and all three integrals in the bottom line are zero by construction. Thus, w is harmonic in $\Omega\backslash\overline{D}$ according to Weyl's Lemma. The same kind of argument also shows that w is harmonic in D. Accordingly, as above, (16) holds true for any $v \in H^1_\diamond(\Omega)$, where now the left-hand side of (16) is zero because of the orthogonality. A standard variational argument then shows that the normal derivative of w on $\partial\Omega$ and the flux of w across ∂D must vanish. ∎

We briefly mention that every potential w from $\mathcal{W}$ of (6) has a unique continuation to a potential $w \in \mathcal{K}$, and the restriction of a nontrivial element from $\mathcal{K}$ to $\Omega\backslash\overline{D}$ is a nonzero element from $\mathcal{W}$. Accordingly, the set of traces on $\partial\Omega$ of potentials from $\mathcal{W}$ and $\mathcal{K}$, respectively, are the same.

Theorem 2. *The operator $K : L^2_\diamond(\partial\Omega) \to \mathcal{X}$ defined in (12) is bounded, injective, and its range lies dense in the subset $\mathcal{K}$ introduced in Lemma 1. The adjoint operator $K^* : \mathcal{X} \to L^2_\diamond(\partial\Omega)$ satisfies*

$$K^* v = \begin{cases} v|_{\partial\Omega}, & v \in \mathcal{K}, \\ 0, & v \in H^1_\diamond(\Omega). \end{cases}$$

In particular, there holds $K^*K = \Lambda_0 - \Lambda_{\mathbb{1}}$, i.e., (11).

Proof. We recall that the two Neumann problems (2) and (5) have well-defined unique solutions u_0 and $u_{\mathbb{1}}$ in the space $H^1_\diamond(\Omega\backslash\overline{D})$ and $H^1_\diamond(\Omega)$, respectively, which are given by the corresponding weak formulations

$$\begin{aligned} \int_{\Omega\backslash\overline{D}} \operatorname{grad} u_0 \cdot \operatorname{grad} v_0 \mathrm{d}x &= \int_{\partial\Omega} f v_0 \,\mathrm{d}s \qquad \text{for every } v_0 \in H^1_\diamond(\Omega\backslash\overline{D}), \\ \int_{\Omega} \operatorname{grad} u_{\mathbb{1}} \cdot \operatorname{grad} v \mathrm{d}x &= \int_{\partial\Omega} f v \,\mathrm{d}s \qquad \text{for every } v \in H^1_\diamond(\Omega). \end{aligned} \tag{17}$$

Moreover, the two solutions depend continuously (in H^1) on the given boundary data $f \in L^2_\diamond(\partial\Omega)$. Accordingly, $w = Kf$ is a well-defined element of $\mathcal{X}$ and K a bounded linear operator from $L^2_\diamond(\partial\Omega)$ to $\mathcal{X}$. The jump condition $\int_{\partial D_i} [w] \mathrm{d}s = 0$ is a consequence of the definition (13) of c_i and the uniqueness of the trace of $u_{\mathbb{1}}$ on ∂D.

Now, choose any $f \in L^2_\diamond(\partial\Omega)$, and denote by u_0 and $u_{\mathbb{1}}$ the corresponding solutions of (2) and (5). As in the definition of Kf, we can extend u_0 to a function

$$\hat{u}_0 = \begin{cases} u_0 & \text{in } \Omega\backslash\overline{D}, \\ c_i & \text{in } \overline{D}_i, i = 1, \ldots, m, \end{cases}$$

in $\mathcal{X}$, such that $Kf = \hat{u}_0 - u_{\mathbb{1}}$. First, for $v \in H^1_\diamond(\Omega)$, we have

$$\begin{aligned} (Kf, v)_{\mathcal{X}} &= (\hat{u}_0, v)_{\mathcal{X}} - (u_{\mathbb{1}}, v)_{\mathcal{X}} \\ &= \int_{\Omega\backslash\overline{D}} \operatorname{grad} u_0 \cdot \operatorname{grad} v \mathrm{d}x - \int_{\Omega} \operatorname{grad} u_{\mathbb{1}} \cdot \operatorname{grad} v \mathrm{d}x = 0 \end{aligned}$$

by virtue of (17), and, hence, $\mathcal{R}(K) \perp H^1_\diamond(\Omega)$. It thus follows from Lemma 1 that $\mathcal{R}(K) \subset \mathcal{K}$ and $\mathcal{N}(K^*) = \overline{\mathcal{R}(K)}^\perp \supset H^1_\diamond(\Omega)$ and, in particular, that $K^*v = 0$ for every $v \in H^1_\diamond(\Omega)$.

Second, for $v \in \mathcal{K}$, we compute

$$(Kf, v)_{\mathcal{X}} = (\hat{u}_0, v)_{\mathcal{X}} - (u_{\mathbb{1}}, v)_{\mathcal{X}} = (\hat{u}_0, v)_{\mathcal{X}},$$

since $u_{\mathbb{1}}$ and v are orthogonal to each other according to Lemma 1. Together with (17) thus follows that

$$(Kf, v)_{\mathcal{X}} = \int_{\Omega\setminus\overline{D}} \operatorname{grad} u_0 \cdot \operatorname{grad} v \mathrm{d}x = \int_{\partial\Omega} fv \,\mathrm{d}s = (f, v)_{L^2(\partial\Omega)},$$

i.e., that $K^*v = v|_{\partial\Omega}$. In particular, for $v = Kf = \hat{u}_0 - u_{\mathbb{1}} \in \mathcal{K}$, we obtain

$$K^*Kf = K^*(\hat{u}_0 - u_{\mathbb{1}}) = (u_0 - u_{\mathbb{1}})|_{\partial D},$$

and, hence, the assertion (11) follows, cf. (4).

Assume now that $\mathcal{R}(K)$ were not dense in $\mathcal{K}$. Then there is some $0 \neq v \in \mathcal{K} \cap \overline{\mathcal{R}(K)}^\perp = \mathcal{K} \cap \mathcal{N}(K^*)$, and since $0 = K^*v = v|_{\partial\Omega}$, this function v has vanishing Dirichlet boundary values on $\partial\Omega$. Moreover, as v belongs to $\mathcal{K}$, it is harmonic in $\Omega\setminus\overline{D}$ with vanishing Neumann boundary values on $\partial\Omega$ (see Lemma 1). Thus, $v^+ = v|_{\Omega\setminus\overline{D}} = 0$ because of the unique solvability of the Cauchy problem for harmonic functions. Using Lemma 1 once more, it follows that $v^- = v|_D$ is also harmonic with vanishing Neumann boundary values on ∂D, and, hence, v^- is constant on each D_i, say $v^-|_{D_i} = v_i^-$, $i = 1, \ldots, m$. Since $\int_{\partial D_i} [v] \mathrm{d}s = -v_i^- |\partial D|$ and as v belongs to $\mathcal{X}$, these constants must all be zero. This is a contradiction to $v \neq 0$, and, hence, $\mathcal{R}(K)$ is dense in $\mathcal{K}$.

Finally, to show injectivity of K, we assume $Kf = 0$ for some $f \in L^2_\diamond(\partial\Omega)$. Then $u_0 = u_{\mathbb{1}}$ in $\Omega\setminus\overline{D}$, and $u_{\mathbb{1}} = c_i$ in D_i, $i = 1, \ldots, m$. Since $u_{\mathbb{1}}$ is harmonic in all of the domain Ω, the field must be constant in Ω (principle of unique continuation), and the flux $f = \partial u_{\mathbb{1}} / \partial \nu$ vanishes on $\partial\Omega$. ■

This theorem – together with Lemma 1 – reveals that the range of K^* consists of all traces of potentials $w \in \mathcal{K}$, whereas the range of $\Lambda_0 - \Lambda_{\mathbb{1}}$ only consists of a dense subset of this set. Accordingly, we need to find a way to deduce the range of K^* from the given data to decrypt the information hidden in these traces according to Theorem 1.

To this end we exploit the so-called *Picard criterion*, a formulation of which can be found in the appendix (Theorem 25) for the ease of completeness. The Picard criterion is based on the singular value decomposition of the operator K, which is largely equivalent to the spectral decomposition of the operator $K^*K = \Lambda_0 - \Lambda_{\mathbb{1}}$.

Corollary 1. *The operator $\Lambda_0 - \Lambda_{\mathbb{1}}$ is a compact and self-adjoint operator from $L^2_\diamond(\partial\Omega)$ into itself. As such, $L^2_\diamond(\partial\Omega)$ has an orthonormal eigenbasis $\{f_j\}$ and*

associated eigenvalues λ_j*, such that*

$$(\Lambda_0 - \Lambda_{\mathbb{1}})\, f_j = \lambda_j\, f_j, \quad n \in \mathbb{N}. \tag{18}$$

These eigenvalues are positive and converge to zero as $n \to \infty$*. Throughout we shall assume that they are sorted in nonincreasing order.*

Proof. That Λ_0 and $\Lambda_{\mathbb{1}}$ are compact operators can be seen from the fact that the trace space of $H^1(\Omega\backslash\overline{D})$ on $\partial\Omega$, i.e., $H^{1/2}(\partial\Omega)$, is compactly embedded in $L^2(\partial\Omega)$. Accordingly, the difference operator $\Lambda_0 - \Lambda_{\mathbb{1}}$ is compact as well as self-adjoint, as follows readily from (11). One can thus find an orthonormal eigenbasis of $\Lambda_0 - \Lambda_{\mathbb{1}}$, and the associated eigenvalues converge to zero for $j \to \infty$. It remains to prove that they are all positive; this follows from (11) and the injectivity of K by Theorem 2. ■

As we have mentioned before, a point $z \in \Omega$ belongs to D, if and only if the trace ϕ_z of U_z is the trace of a potential in $\mathcal{K}$, i.e., if it belongs to the range of K^*. As we show in the appendix, cf. Corollary 3, this can be tested in the following way.

Theorem 3. *Let* $\{f_j\}$ *and* $\{\lambda_j\}$ *be the eigenbasis and eigenvalues of* $\Lambda_0 - \Lambda_{\mathbb{1}}$*. Then, for any point* $z \in \Omega$*,*

$$z \in D \iff \sum_{n=1}^{\infty} \frac{\left|(\phi_z, f_j)_{L^2(\partial\Omega)}\right|^2}{\lambda_j} < \infty \tag{19}$$

with $\phi_z = U_z|_{\partial\Omega}$ *from (8).*

Remark 1. With the notations $1/\infty = 0$ and $\operatorname{sign} \alpha = \begin{cases} \alpha/|\alpha|, & \alpha \neq 0, \\ 0, & \alpha = 0, \end{cases}$ for any $\alpha \in \mathbb{C}$, we note that

$$\mathcal{X}_D(z) = \operatorname{sign}\left[\sum_j \frac{\left|(\phi_z, f_j)_{L^2(\partial\Omega)}\right|^2}{\lambda_j}\right]^{-1}, \quad z \in \Omega,$$

is the characteristic function of D. In particular, this result provides a constructive proof of the uniqueness of the inverse problem.

Conducting Obstacles

Next, we turn to the case of anisotropic conducting obstacles. To this end we assume that for each $x \in \Omega$ the conductivity $\sigma(x)$ is a real, symmetric positive definite $n \times n$-matrix, measurable and essentially bounded as a function of x and that the associated quadratic form is bounded from below by some positive constant $c > 0$,

i.e.,

$$p \cdot (\sigma(x)p) \geq c \text{ for almost every } x \in \overline{D} \text{ and every } p \in \mathbb{R}^n \text{ with } |p| = 1 \text{ and } \sigma(x) = I \text{ on } \Omega \backslash \overline{D}, \tag{20}$$

where D denotes the obstacles, which are assumed to have the same topological properties as before. Another assumption that seems to be necessary for the validity of the factorization method is that

$$p \cdot (\sigma(x)p) \leq \kappa < 1 \quad \text{for every } p \in \mathbb{R}^n \text{ with } |p| = 1, \text{ and almost every } x \in D, \tag{21}$$

which states that the background conductivity of the object is strictly larger than within the inclusions. Instead of (21), one can alternatively require that the conductivity within the inclusions is strictly larger than in the background, with straightforward modifications of the analysis; however, we will stick to the above assumption for the ease of simplicity. We mention that the assumption that the background conductivity be *strictly* larger (or smaller) than within the object can be relaxed to just being larger (or smaller), for the prize that the outcome of the method, is unspecified for sampling points right on the boundary of the inclusions, cf. [42]. However, it is an open problem whether the factorization method is applicable, if inequality (21) holds in some obstacles, while $p \cdot (\sigma(x)p) \geq \gamma > 1$ in other inclusions; numerically, the method does not seem to deterior in this "mixed case."

With conducting obstacles, the potential corresponding to a boundary current $f \in L^2_\diamond(\partial\Omega)$ is given as the (weak) solution $u \in H^1_\diamond(\Omega)$ of the boundary value problem

$$\operatorname{div}(\sigma \operatorname{grad} u) = 0 \quad \text{in } \Omega, \qquad \frac{\partial}{\partial \nu} u = f \quad \text{on } \partial\Omega, \qquad \int_{\partial\Omega} u \, \mathrm{d}s = 0, \tag{22}$$

which replaces the model (2) from section "Impedance Tomography in the Presence of Insulating Inclusions" above. Accordingly, we denote by Λ the Neumann-Dirichlet map associated with (13), i.e., $\Lambda : f \mapsto g = u|_{\partial\Omega}$.

As before, the corresponding **inverse problem** is to determine the shape of the obstacles D from the relative data $\Lambda - \Lambda_{\mathbb{1}}$. Here, again, $\Lambda_{\mathbb{1}}$ corresponds to the "unperturbed" case $\sigma = \sigma_{\mathbb{1}} = 1$ everywhere in Ω.

We mention that the problem whether not only D but the conductivity σ itself is uniquely determined by these data is completely settled when $n = 2$ – as long as σ is isotropic – cf. [14]. For $n = 3$, this question is still open for general scalar L^∞ – conductivities. Partial answers are known; we refer to chapter ▸ Electrical Impedance Tomography. However, the set D is uniquely determined as we will see below in Theorem 7.

Now we proceed to derive a factorization of $\Lambda - \Lambda_{\mathbb{1}}$ in three factors as in (1), i.e.,

$$\Lambda - \Lambda_{\mathbb{1}} = AGA^*. \tag{23}$$

To this end we imagine the effect of a *virtual source* φ on the boundary of the obstacle D, given that the boundary of the object Ω is insulated. The corresponding potential v is the solution of the boundary value problem

$$\begin{aligned} \Delta v = 0 \quad &\text{in } \Omega\backslash\overline{D}, \qquad -\frac{\partial}{\partial \nu} v = \varphi \quad \text{on } \partial D, \\ &\frac{\partial}{\partial \nu} v = 0 \quad \text{on } \partial\Omega, \qquad \int_{\partial\Omega} v \, \mathrm{d}s = 0. \end{aligned} \tag{24}$$

Recall that the normal vector ν on ∂D has been fixed to point into the interior of $\Omega\backslash\overline{D}$, and therefore the minus sign in front of the normal derivative on ∂D reflects the fact that φ is considered to be a source and not a sink. We will require that this source has vanishing mean on *each* connected component D_i of D, i.e.,

$$\varphi \in H_*^{-1/2}(\partial D) = \left\{ \varphi \in H^{-1/2}(\partial D) : \int_{\partial D_i} \varphi \mathrm{d}s = 0, i = 1, \ldots, m \right\}, \tag{25}$$

where the integrals have to be interpreted as dual pairings between $H^{-1/2}$ functions and the unit constant from $H^{1/2}$. For later use we remark that the dual space of $H_*^{-1/2}(\partial D)$ can be identified with the subspace

$$H_*^{1/2}(\partial D) = \left\{ \psi \in H^{1/2}(\partial D) : \int_{\partial D_i} \psi \mathrm{d}s = 0, i = 1, \ldots, m \right\} \tag{26}$$

of $H^{1/2}(\partial D)$.

Associated with (24), we define the operator

$$A : \begin{cases} H_*^{-1/2}(\partial D) \to L^2_\diamond(\partial\Omega), \\ \varphi \quad \mapsto \quad v|_{\partial\Omega}, \end{cases} \tag{27}$$

and remark that the adjoint operator $A^* : L^2_\diamond(\partial\Omega) \to H_*^{1/2}(\partial D)$ of A is easily seen to map $f \in L^2_\diamond(\partial\Omega)$ onto the trace of the solution u_0 of (2) on the boundary of the obstacle – after an appropriate renormalization of this trace on each component ∂D_i of ∂D. More precisely the following holds

$$(A^* f)(x) = u_0(x) - c_i \quad \text{for } x \in \partial D_i, \ i = 1, \ldots, m, \tag{28}$$

with c_i as in (13).

In order to establish (23), it remains to determine the operator G in the middle. We define G via the weak solution w of the diffraction problem

$$\begin{aligned} \operatorname{div}(\sigma \operatorname{grad} w) = 0 \quad \text{in } \Omega\backslash\partial D, \qquad & \frac{\partial}{\partial \nu} w = 0 \quad \text{on } \partial\Omega, \qquad \int_{\partial\Omega} w \mathrm{d}s = 0, \\ [w]_{\partial D} = \psi, \qquad & [\nu \cdot (\sigma \operatorname{grad} w)]_{\partial D} = 0, \end{aligned} \tag{29}$$

and the solution $w_{\mathbb{1}}$ of the corresponding problem with σ replaced by one everywhere. Again, the normal ν on ∂D is pointing into the exterior of D. Note that when $\sigma = 1$ throughout all of Ω, then the corresponding solution $w_{\mathbb{1}}$ of (29) can be represented as a modified double-layer potential with density ψ and the Neumann function for the Laplacian as kernel, i.e.,

$$w_{\mathbb{1}}(x) = \int_{\partial D} \frac{\partial}{\partial_y \nu} N(x, y) \psi(y) \mathrm{d}s(y), \qquad x \in \Omega\backslash\partial D.$$

For a general conductivity tensor, the weak form of (29) is obtained by integrating the differential equation against any test function $v \in H^1(\Omega)$ and using partial integration, which yields

$$\int_{\Omega\backslash\partial D} \operatorname{grad} w \cdot (\sigma \operatorname{grad} v) \, \mathrm{d}x = 0 \qquad \text{for every } v \in H^1(\Omega). \tag{30}$$

Now we can make the ansatz $w = w_{\mathbb{1}} + \hat{w}$ with $\hat{w} \in H^1(\Omega)$ to rewrite this as a standard variational problem in $H^1(\Omega)$. Find $\hat{w} \in H^1(\Omega)$ such that

$$\int_{\Omega} \operatorname{grad} \hat{w} \cdot (\sigma \operatorname{grad} v) \, \mathrm{d}x = -\int_{\Omega\backslash\partial D} \operatorname{grad} w_{\mathbb{1}} \cdot (\sigma \operatorname{grad} v) \mathrm{d}x$$

for every $v \in H^1(\Omega)$. From this, it follows readily that problem (29) has a unique weak solution in $H^1(\Omega\backslash\partial D)$, provided that $\psi \in H^{1/2}(\partial D)$, i.e., that ψ belongs to the trace space of $H^1(D)$. In accordance with the definition of A^*, however, we will restrict ψ to $H_*^{1/2}(\partial D)$.

The flux of w and $w_{\mathbb{1}}$ across ∂D is well defined in $H^{-1/2}(\partial D)$, cf., e.g., [45, Thm. 2.5], and there holds

$$\begin{aligned} \int_{\partial D_i} \frac{\partial}{\partial \nu} (w^+ - w_{\mathbb{1}}^+) \, \mathrm{d}s &= \int_{\partial D_i} \nu \cdot (\sigma \operatorname{grad} w^-) \mathrm{d}s - \int_{\partial D_i} \frac{\partial}{\partial \nu} w_{\mathbb{1}}^- \mathrm{d}s \\ &= \int_{D_i} \operatorname{div}(\sigma \operatorname{grad} w) \mathrm{d}x - \int_{D_i} \Delta w_{\mathbb{1}} \mathrm{d}x = 0. \end{aligned}$$

We can therefore define the bounded operator G in the following way:

$$G : \begin{cases} H_*^{1/2}(\partial D) & \to \quad H_*^{-1/2}(\partial D), \\ \psi & \mapsto \quad \dfrac{\partial}{\partial \nu} (w^+ - w_{\mathbb{1}}^+)|_{\partial D}. \end{cases} \tag{31}$$

Theorem 4. *With A and G defined as above, the difference $\Lambda - \Lambda_{\mathbb{1}}$ of the two Neumann-Dirichlet operators associated with (22) and (5), respectively, satisfies*

$$\Lambda - \Lambda_{\mathbb{1}} = AGA^*.$$

Proof. Consider an arbitrary element $f \in L^2_\diamond(\partial\Omega)$ and the corresponding function $\psi = A^* f$, which satisfies

$$\psi|_{\partial D_i} = u_0|_{\partial D_i} - c_i,$$

where u_0 is given by (2), and c_i is as in (13). The function ψ belongs to $H_*^{1/2}(\partial D)$, and it is easy to verify that the associated solution $w_{\mathbb{1}}$ of (29) – where σ is replaced by one – is given by

$$w_{\mathbb{1}} = \begin{cases} u_0 - u_{\mathbb{1}} & \text{in } \Omega\backslash\overline{D}, \\ c_i - u_{\mathbb{1}} & \text{in } D_i, i = 1, \ldots, m, \end{cases}$$

where $u_{\mathbb{1}}$ is the solution of (5). Similarly, the solution w of (29) is given by

$$w = \begin{cases} u_0 - u & \text{in } \Omega\backslash\overline{D}, \\ c_i - u & \text{in } D_i, i = 1, \ldots, m, \end{cases}$$

with u from (22). Accordingly, $w^+ - w_{\mathbb{1}}^+ = u_{\mathbb{1}}^+ - u^+$, and hence,

$$\varphi = GA^* f = \frac{\partial}{\partial \nu}(u_{\mathbb{1}}^+ - u^+)\Big|_{\partial D}.$$

If we insert this particular source term φ into (24), then we conclude readily that the associated solution v of (24) is given by $v = u^+ - u_{\mathbb{1}}^+$. It thus follows from (27) that

$$AGA^* f = A\varphi = g - g_{\mathbb{1}} = (\Lambda - \Lambda_{\mathbb{1}}) f$$

as required. ∎

At this occasion, we recall that every function $w \in \mathcal{W}$ of (6) has a well-defined normal derivative $\varphi \in H_*^{-1/2}(\partial D)$ at the inner boundary ∂D and, hence, solves the corresponding boundary value problem (24). And vice versa, the solution of (24) for any $\varphi \in H_*^{-1/2}(\partial D)$ belongs to $\mathcal{W}$. Thus, we can reformulate Theorem 1 as follows.

Theorem 5. *A point $z \in \Omega$ belongs to D, if and only if the trace ϕ_z of the dipole potential U_z in z, defined by (8), belongs to $\mathcal{R}(A)$.*

As in the insulating case, it remains to derive a constructive algorithm to test whether the trace of some dipole potential belongs to $\mathcal{R}(A)$ or not. The next step on

our way towards this goal is an investigation of the functional analytic properties of the operator G. In the following, we will often consider operators acting between a reflexive Banach space X and its dual space X^*. We will denote the action of an element $\ell \in X^*$ on an element $\psi \in X$ by $\langle \ell,\ \psi \rangle$ and the pair of spaces by $\langle X^*,\ X \rangle$ in order to indicate that the first argument belongs to X^* and the second to X. A particular example is the Sobolev space $H_*^{1/2}(\partial D)$ with dual space $H_*^{-1/2}(\partial D)$.

Theorem 6. *The operator $G : H_*^{1/2}(\partial D) \to H_*^{-1/2}(\partial D)$ is self-adjoint (i.e., G coincides with $G^* : H_*^{1/2}(\partial D) \to H_*^{-1/2}(\partial D)$ if the bi-dual of $H_*^{1/2}(\partial D)$ is identified with itself) and coercive, i.e., there exists $\gamma > 0$ with*

$$\langle G\psi, \psi \rangle \geq \gamma \, \|\psi\|^2_{H^{1/2}(\partial D)} \quad \text{for all } \psi \in H_*^{1/2}(\partial D). \tag{32}$$

Here, $\langle \cdot,\ \cdot \rangle$ denotes the dual pairing in the dual system $\left\langle H_*^{-1/2}(\partial D), H_*^{1/2}(\partial D) \right\rangle$.

Proof. The proof proceeds in a couple of steps.

1. At first we establish the symmetry of G. Take any ψ and $\tilde{\psi}$ from $H_*^{1/2}(\partial D)$, define w and $w_{\mathbb{1}}$ as in the proof of Theorem 4, and – using $\tilde{\psi}$ instead of ψ in (29) – define $\tilde{w}$ and $\tilde{w}_{\mathbb{1}}$ accordingly. Then we conclude that

$$\begin{aligned}
\langle G\psi, \tilde{\psi} \rangle &= \int_{\partial D} \tilde{\psi} \frac{\partial}{\partial \nu} w^+ \mathrm{d}s - \int_{\partial D} \tilde{\psi} \frac{\partial}{\partial \nu} w_{\mathbb{1}}^+ \mathrm{d}s \\
&= \int_{\partial D} (\tilde{w}^+ - \tilde{w}^-) \frac{\partial}{\partial \nu} w^+ \mathrm{d}s - \int_{\partial D} \left(\tilde{w}_{\mathbb{1}}^+ - \tilde{w}_{\mathbb{1}}^-\right) \frac{\partial}{\partial \nu} w_{\mathbb{1}}^+ \mathrm{d}s \\
&= \int_{\partial D} \tilde{w}^+ \frac{\partial}{\partial \nu} w^+ \mathrm{d}s - \int_{\partial D} \tilde{w}^- (\nu \cdot (\sigma \operatorname{grad} w^-)) \mathrm{d}s \\
&\quad - \int_{\partial D} \tilde{w}_{\mathbb{1}}^+ \frac{\partial}{\partial \nu} w_{\mathbb{1}}^+ \mathrm{d}s + \int_{\partial D} \tilde{w}_{\mathbb{1}}^- \frac{\partial}{\partial \nu} w_{\mathbb{1}}^- \mathrm{d}s.
\end{aligned}$$

Now we can use (29) and apply Green's formula in D or $\Omega \backslash \overline{D}$, respectively, in each of these integrals (care has to be taken concerning the orientation of the normal on ∂D), to obtain

$$\begin{aligned}
\langle G\psi, \tilde{\psi} \rangle &= -\int_{\Omega \backslash \overline{D}} \operatorname{grad} \tilde{w} \cdot \operatorname{grad} w \mathrm{d}x - \int_D \operatorname{grad} \tilde{w} \cdot (\sigma \operatorname{grad} w)\ \mathrm{d}x \\
&\quad + \int_{\Omega \backslash \overline{D}} \operatorname{grad} \tilde{w}_{\mathbb{1}} \cdot \operatorname{grad} w_{\mathbb{1}} \mathrm{d}x + \int_D \operatorname{grad} \tilde{w}_{\mathbb{1}} \cdot \operatorname{grad} w_{\mathbb{1}} \mathrm{d}x \\
&= \int_{\Omega \backslash \partial D} \operatorname{grad} \tilde{w}_{\mathbb{1}} \cdot \operatorname{grad} w_{\mathbb{1}} \mathrm{d}x - \int_{\Omega \backslash \partial D} \operatorname{grad} \tilde{w} \cdot (\sigma \operatorname{grad} w)\ \mathrm{d}x,
\end{aligned} \tag{33}$$

from which the symmetry of G is obvious.

2. Turning to the coercivity assertion (32), we fix some $\psi \in H_*^{1/2}(\partial D)$ and employ the weak form (30) of (29) with $v = w - w_{\mathbb{1}} \in H^1(\Omega)$. Starting from (33) with $\psi = \tilde{\psi}$, we thus obtain

$$\begin{aligned}
\langle G\psi, \psi \rangle &= \int_{\Omega \setminus \partial D} |\operatorname{grad} w_{\mathbb{1}}|^2 \mathrm{d}x - \int_{\Omega \setminus \partial D} \operatorname{grad} w \cdot (\sigma \operatorname{grad} w)\, \mathrm{d}x \\
&= \int_{\Omega \setminus \partial D} |\operatorname{grad} w_{\mathbb{1}}|^2 \mathrm{d}x - \int_{\Omega \setminus \partial D} \operatorname{grad} w \cdot (\sigma \operatorname{grad} w)\, \mathrm{d}x \\
&\quad + 2 \int_{\Omega \setminus \partial D} \operatorname{grad} w \cdot (\sigma \operatorname{grad} (w - w_{\mathbb{1}}))\, \mathrm{d}x \\
&= \int_{\Omega \setminus \partial D} |\operatorname{grad} w_{\mathbb{1}}|^2 \mathrm{d}x + \int_{\Omega \setminus \partial D} \operatorname{grad} w \cdot (\sigma \operatorname{grad} w)\, \mathrm{d}x \\
&\quad - 2 \int_{\Omega \setminus \partial D} \operatorname{grad} w \cdot (\sigma \operatorname{grad} w_{\mathbb{1}})\, \mathrm{d}x \\
&= \int_{\Omega \setminus \partial D} \operatorname{grad} w_{\mathbb{1}} \cdot ((1 - \sigma) \operatorname{grad} w_{\mathbb{1}})\, \mathrm{d}x \\
&\quad + \int_{\Omega \setminus \partial D} \operatorname{grad} (w - w_{\mathbb{1}}) \cdot (\sigma \operatorname{grad} (w - w_{\mathbb{1}}))\, \mathrm{d}x \\
&\geq \int_{\Omega \setminus \partial D} \operatorname{grad} w_{\mathbb{1}} \cdot ((1 - \sigma) \operatorname{grad} w_{\mathbb{1}})\, \mathrm{d}x.
\end{aligned}$$

The integrand of the last integral vanishes in $\Omega \setminus \overline{D}$ and can be bounded in D from below using the restriction (21) on the conductivity. Accordingly we have

$$\langle G\psi, \psi \rangle \geq (1 - \kappa) \int_D |\operatorname{grad} w_{\mathbb{1}}|^2 \mathrm{d}x. \tag{34}$$

3. To accomplish the proof of (32), we need to show that

$$\|\operatorname{grad} w_{\mathbb{1}}\|_{L^2(D)} \geq c\, \|\psi\|_{H^{1/2}(\partial D)} \tag{35}$$

for some constant $c > 0$. Assume the contrary: let $\psi^{(j)} \in H_*^{1/2}(\partial D)$ and the corresponding $w_{\mathbb{1}}^{(j)}$ be such that $\|\psi^{(j)}\|_{H^{1/2}(\partial D)} = 1$ for every j and that $\left\|\operatorname{grad} w_{\mathbb{1}}^{(j)}\right\|_{L^2(D)}$ converges to zero as j tends to infinity. Define $\tilde{w}_{\mathbb{1}}^{(j)} \in H^1(\Omega \setminus \partial D)$ as

$$\tilde{w}_{\mathbb{1}}^{(j)} = \begin{cases} w_{\mathbb{1}}^{(j)} & \text{in } \Omega \setminus D, \\ w_{\mathbb{1}}^{(j)} - c_i^{(j)} & \text{in } D_i, i = 1, \ldots, m, \end{cases}$$

with

$$c_i^{(j)} = \frac{1}{|\partial D_i|} \int_{\partial D_i} \left(w_{\mathbb{1}}^{(j)}\right)^- \mathrm{d}s, \qquad i = 1, \ldots, m.$$

Then $\tilde{w}_{\mathbb{1}}^{(j)}|_{D_i}$ has vanishing mean on ∂D_i, and $\left\|\operatorname{grad} \tilde{w}_{\mathbb{1}}^{(j)}\right\|_{L^2(D_i)} \to 0$ for every $i = 1, \ldots, m$ as $j \to \infty$. By virtue of the Poincaré inequality, this implies that $\tilde{w}_{\mathbb{1}}^{(j)}$ tends to zero in $H^1(D)$. From (29), thus follows that the normal derivative $\frac{\partial}{\partial \nu} \tilde{w}_{\mathbb{1}}^{(j)}$ at ∂D (from either side) tends to zero in $H^{-1/2}(\partial D)$ and, hence, that $\tilde{w}_{\mathbb{1}}^{(j)}|_{\Omega \setminus \overline{D}}$ converges in $H^1(\Omega \setminus \overline{D})$ to the solution of the homogeneous Neumann problem, normalized at the outer boundary. In other words, $\tilde{w}_{\mathbb{1}}^{(j)}$ converges to zero in $H^1(D)$ and in $H^1(\Omega \setminus \overline{D})$ as $j \to \infty$. Recurring to (29) once again, we observe that

$$\psi^{(j)}|_{\partial D_i} + c_i^{(j)} = \left[w_{\mathbb{1}}^{(j)}\right]_{\partial D_i} + c_i^{(j)} = \left[\tilde{w}_{\mathbb{1}}^{(j)}\right]_{\partial D_i}, \tag{36}$$

and since $\psi^{(j)} \in H_*^{1/2}(\partial D)$, it follows by integration over ∂D_i that

$$\begin{aligned} c_i^{(j)} &= \frac{1}{|\partial D_i|} \int_{\partial D_i} \left[\tilde{w}_{\mathbb{1}}^{(j)}\right]_{\partial D_i} \mathrm{d}s - \frac{1}{|\partial D_i|} \int_{\partial D_i} \psi^{(j)} \mathrm{d}s \\ &= \frac{1}{|\partial D_i|} \int_{\partial D_i} \left[\tilde{w}_{\mathbb{1}}^{(j)}\right]_{\partial D_i} \mathrm{d}s \to 0 \end{aligned}$$

as j runs to infinity. Inserting this into (36), we conclude that

$$\psi^{(j)}|_{\partial D_i} = \left[\tilde{w}_{\mathbb{1}}^{(j)}\right]_{\partial D_i} - c_i^{(j)} \to 0, \quad j \to \infty$$

in $H^{1/2}(\partial D_i)$, $i = 1, \ldots, m$, but this contradicts $\left\|\psi^{(j)}\right\|_{H^{1/2}(\partial D)} = 1$. Therefore, (35) is true for some $c > 0$ and every $\psi \in H_*^{1/2}(\partial D)$, and hence, (32) follows from (34) and (35). ∎

By virtue of Theorem 6, all assumptions of Corollary 5 are satisfied for the factorization of the relative data $\Lambda - \Lambda_{\mathbb{1}}$ established in Theorem 5. Therefore, we can now conclude the main result of this section.

Theorem 7. *Let $z \in \Omega$ and ϕ_z be defined as before.*

Then:
$$z \in D \iff \sum_{n=1}^{\infty} \frac{\left|(\phi_z, f_j)_{L^2(\partial\Omega)}\right|^2}{\lambda_j} < \infty,$$

where f_j and λ_j are the orthonormal eigenfunctions and eigenvalues of $\Lambda - \Lambda_{\mathbb{1}}$.

Local Data

It is an important feature of the factorization method that it can be easily adapted to applications where the given data correspond to what is called the local Neumann-Dirichlet map Λ^ℓ. This is the map that takes Neumann boundary values supported on some relatively open subset $\Gamma \subset \partial\Omega$ only and returns the corresponding boundary potentials on the very same subset (normalized to have vanishing mean, say). The local Neumann-Dirichlet map occurs whenever part of the boundary is inaccessible to measurements, in which case Γ corresponds to that part of the boundary of Ω where electrodes can be attached. Mathematically, the local Neumann-Dirichlet map can be interpreted as a Galerkin projection

$$\Lambda^\ell = P\Lambda P^* \tag{37}$$

of the full Neumann-Dirichlet map, where

$$P : \begin{cases} L^2_\diamond(\partial\Omega) & \to & L^2_\diamond(\Gamma), \\ g & \mapsto & g|_\Gamma - \frac{1}{|\Gamma|}\int_\Gamma g\,\mathrm{d}s, \end{cases} \tag{38}$$

and P^* is its L^2 adjoint, i.e.,

$$P^* f = \begin{cases} f & \text{on } \Gamma, \\ 0 & \text{on } \partial D \setminus \Gamma. \end{cases}$$

From Theorem 4 we immediately conclude that if the conductivity distribution satisfies (20) and (21), then the difference of the two local Neumann-Dirichlet maps Λ^ℓ and $\Lambda^\ell_{\mathbb{1}}$ can be factorized in the form

$$\Lambda^\ell - \Lambda^\ell_{\mathbb{1}} = (PA)G(PA)^*$$

with A and G as before. Moreover, the coercivity of G allows a constructive way to check whether a given function belongs to $\mathcal{R}(PA)$, considered as an operator from $H^{-1/2}_*(\partial D)$ to $L^2_\diamond(\Gamma)$. Note that it is obvious from Theorem 5 that the function $P\phi_z$ belongs to $\mathcal{R}(PA)$ when $z \in D$; the converse statement requires a little more effort.

Theorem 8. *Let Γ be a relatively open subset of $\partial\Omega$, and let P be the projector defined in (38). Then $z \in D$, if and only if $P\phi_z \in \mathcal{R}(PA)$.*

Proof. According to the definition (27) of A, the test function $P\phi_z$ belongs to $\mathcal{R}(PA)$, if and only if ϕ_z coincides on Γ (up to a constant) with the trace of a solution v of (24). In this case, however, the dipole potential U_z and the function v are both harmonic functions in $\Omega\setminus(\overline{D} \cup \{z\})$ and have the same Cauchy data on Γ (again, up to a constant). Now we choose a connected subset Ω' of $\Omega\setminus(\overline{D} \cup \{z\})$, whose boundary contains a portion of Γ that is also a relatively open subset of $\partial\Omega$. Then U_z

and v coincide up to a constant in Ω' according to Holmgren's theorem and, hence, near all of $\partial\Omega$. This shows that $\phi_z \in \mathcal{R}(A)$, and, hence, the assertion follows from Theorem 5. ■

Accordingly, if Γ is a relatively open subset of $\partial\Omega$, then Theorem 7 also extends readily to the local situation, if the eigenfunctions and eigenvalues of $\Lambda - \Lambda_{\mathbb{1}}$ are replaced by those of $\Lambda^{\ell} - \Lambda_{\mathbb{1}}^{\ell}$.

Note that Theorem 8 requires that Γ is a relatively open subset of $\partial\Omega$, and in fact, the factorization method no longer applies for discrete measurements or finitely many boundary currents. Still, this is precisely the situation that is encountered in practice, as data are always finite dimensional. Due to the rapid decay of the eigenvalues of $\Lambda - \Lambda_{\mathbb{1}}$, however, the full relative data can be very well approximated by operators of finite rank, such as those corresponding to real data (see [54] for detailed numerical examples).

Other Generalizations

The Half-Space Problem

The factorization method can also be applied to a related inverse electrostatic problem in full space with near field data, if the same manifold of codimension one is used to generate a source *and* to measure the resulting change of the potential. In fact, this problem which has been studied in [52] and [75] is very similar to the setting for the Helmholtz equation that we will consider in the following section. We also like to refer to [15] where this approach has been applied to some real two-dimensional data.

For quite a few applications, however, the impedance tomography problem is more appropriately modeled in a half space, rather than in the full space or within a bounded domain. For this setting new difficulties arise, as the data (may) live on the entire, unbounded boundary of the surface, which calls for weighted Sobolev spaces for an appropriate theoretical analysis. In the sequel we restrict our attention to three-space dimensions ($n = 3$), as the two-dimensional case needs some additional attention, cf. [55] and at the same time appears to be less interesting from a practical point of view.

We consider the half-space $\Omega = \{x \in \mathbb{R}^3 : \nu \cdot x < 0\}$, where $\nu \in \mathbb{R}^3$ is a fixed unit vector, which coincides with the outer normal on the hyperplane $\{x : \nu \cdot x = 0\}$, which is the boundary of Ω. The main difficulty in the analysis of this problem is that solutions of the corresponding conductivity problem

$$\operatorname{div}(\sigma \operatorname{grad} u) = 0 \quad \text{in } \Omega, \qquad \frac{\partial}{\partial \nu} u = f \quad \text{on } \partial\Omega, \tag{39}$$

need no longer belong to $L^2(\Omega)$; instead, one has to resort to weighted Sobolev spaces, such as

$$\mathcal{U} = \{u \in \mathcal{D}'(\Omega) : (1 + |\cdot|^2)^{-1/2} u \in L^2(\Omega), |\operatorname{grad} u| \in L^2(\Omega)\},$$

to search for a unique solution of (39). If σ is given by (20), then a weak solution $u \in \mathcal{U}$ can be shown to exist provided that f belongs to

$$L^{2,-1}(\partial\Omega) = \{f : (1 + |\cdot|^2)^{1/2} f \in L^2(\partial\Omega)\},$$

in which case the trace of u belongs to the dual space $L^{2,1}(\partial\Omega)$ of $L^{2,-1}(\partial\Omega)$. Note that no normalization of u is required in (39) because solutions in $\mathcal{U}$ are implicitly normalized to vanish at infinity. We refer to [55] for further details about the forward problem.

Within this function space setting, the Neumann-Dirichlet operator is defined in a natural way as an operator $\Lambda : L^{2,-1}(\partial\Omega) \to L^{2,1}(\partial\Omega)$, and the difference between Λ and $\Lambda_{\mathbb{1}}$ (the latter corresponding to the homogeneous half space) admits a factorization (23) as before, where now

$$A : \begin{cases} H_*^{-1/2}(\partial D) & \to & L^{2,1}(\partial\Omega), \\ \varphi & \mapsto & v|_{\partial\Omega}, \end{cases}$$

and v solves the same boundary value problem as in (24), except for the missing normalization over the boundary $\partial\Omega$. Furthermore, the self-adjoint operator G is defined as before (with the appropriate definition of a weak solution of (29)) and is coercive again.

We emphasize that the dipole potential (8) for the half space is explicitly known, i.e., we have (up to a negligible multiplicative constant)

$$\phi_z(x) = \frac{(x - z) \cdot p}{|x - z|^3}, \qquad x \in \partial\Omega. \tag{40}$$

With these notations, the characterization of the inclusions can be established in much the same way as before (see [55]).

Theorem 9. *A point $z \in \Omega$ belongs to D, if and only if ϕ_z of (40) belongs to $\mathcal{R}(A)$.*

For real applications, the measuring device will only cover a bounded region $\Gamma \subset \partial\Omega$. The corresponding local Neumann-Dirichlet operator Λ^ℓ can then be embedded in the standard L^2 framework from the previous section, and the usual Picard series can be used to implement the range test. For the ease of completeness, we briefly mention that for such local data the test dipole ϕ_z can be replaced by the function

$$\tilde{\phi}_z(x) = \frac{1}{|x - z|}, \qquad x \in \Gamma,$$

which is the trace of the corresponding Neumann function (again, up to a multiplicative constant), as the latter has a vanishing normal derivative on the boundary of the half space. We hasten to add, though, that $\tilde{\phi}_z$ must not be used for full data, as it does not belong to $L^{2,1}(\partial\Omega)$. Numerically, however, this modification of the method has no significant benefit.

The Crack Problem

Another case of interest are cracks, i.e., lower-dimensional manifolds of codimension one, that are insulating, say. This setting has important applications in nondestructive testing of materials. Consider a domain $\Omega \subset \mathbb{R}^n$, with $n = 2$ or $n = 3$ again, and the union $\Sigma = \cup_{i=1}^m \Sigma_i \subset \Omega$ of m smooth, bounded manifolds (the insulating cracks), such that $\Sigma_i \cap \Sigma_j = \emptyset$ and $\Omega\backslash\Sigma$ are connected. Given a boundary current $f \in L^2_\diamond(\partial\Omega)$, the induced potential satisfies the model equations

$$\Delta u_0 = 0 \quad \text{in } \Omega\backslash\Sigma, \qquad \frac{\partial}{\partial\nu} u_0 = 0 \quad \text{on } \Sigma, \qquad \frac{\partial}{\partial\nu} u_0 = f \quad \text{on } \partial\Omega, \tag{41}$$

and the corresponding Neumann–Dirichlet operator is the map that takes f onto the trace of u_0 on $\partial\Omega$:

$$\Lambda : \begin{cases} L^2_\diamond(\partial\Omega) & \to & L^2_\diamond(\partial\Omega), \\ f & \mapsto & u_0|_{\partial\Omega}. \end{cases}$$

The crack case can be analyzed in a similar way as in section "Impedance Tomography in the Presence of Insulating Inclusions", cf. [23], using a factorization $\Lambda - \Lambda_{\mathbb{1}} = K^*K$, where K is almost identical to the operator in (12), except that it maps into $H^1(\Omega\backslash\Sigma)$. There is a more important difference, though. As the crack has no interior points, the range test will always fail with the hitherto used test function ϕ_z, as the dipole singularity of U_z is too strong to belong to $H^1(\Omega\backslash\Sigma)$, even when $z \in \Sigma$. To detect a crack, we therefore need to construct a new test function by integrating the function ϕ_z over z along some "test arc" (in $\mathbb{R}^2$) or some "test surface" (in $\mathbb{R}^3$).

The range test can then be implemented by placing linear (planar) test cracks in different sampling points with various orientations (see [23] for numerical reconstructions in two-space dimensions). The amount of work thus grows substantially, as we now have 2 degrees of freedom to sample (a test point and a normal direction) instead of only one in the previous cases. Also, in a numerical realization, test cracks will – at best – only touch the crack tangentially, but in theory this already suffices to ruin the range test. It turns out that in practice the usual implementation with the test function ϕ_z performs as good as the more elaborate but expensive variant described above. As said before, in theory, ϕ_z will never belong to the range of K; in practice, however, it will "almost" do so, i.e., the Picard series (19) will grow much more slowly in the close neighborhood of the crack.

One-dimensional cracks in three-dimensional objects cannot be reconstructed in this way, because the potential does not "see" inhomogeneities of this size.

However, one can use an asymptotic analysis similar to the derivation of MUSIC-type algorithms that are discussed in section "MUSIC" below. Here we give a brief sketch of an argument provided in [47] and refer to this paper for further details. The basic idea is that realistic "one-dimensional" cracks in a 3D world are not exactly one-dimensional, but better modeled as extremely thin tubular inclusions of small diameter $\delta > 0$. The corresponding relative data $\Lambda_\delta - \Lambda_{\mathbb{1}}$, where Λ_δ is the Neumann-Dirichlet operator associated with the tubular inclusion and $\Lambda_{\mathbb{1}}$ is as usual, turn out to satisfy an asymptotic expansion in δ,

$$\Lambda_\delta - \Lambda_{\mathbb{1}} = \delta^2 \hat{M} + o(\delta^2),$$

possibly after selecting an appropriate (sub)sequence $\delta_k \to 0$. The operator $\hat{M}$ that constitutes the dominating term of this expansion admits a factorization similar to (23). In contrast to the MUSIC framework below, this operator has infinite dimensional range. Although the operators of the corresponding factorization are somewhat different from the ones that we have encountered above, the bottom line is the same as for one-dimensional cracks in two-space dimensions. The same integrated test function belongs to the range of the operator A of this factorization, if and only if the corresponding test arc is part of the crack. The singular value decomposition of $\hat{M}$ can be used to evaluate this test, and in practice this singular value decomposition can be approximated by the one of $\Lambda_\delta - \Lambda_{\mathbb{1}}$, i.e., by the given data.

3 The Factorization Method in Inverse Scattering Theory

The second part of this chapter is devoted to the factorization method for problems in inverse scattering theory for time-harmonic waves. The scattering of an incident plane wave by a medium gives rise to a scattered field which is measured "far away" from the medium. The factorization method characterizes the shape of the scattering medium from this far field information. The measurement operator will be the far field operator F which maps the density of the incident Herglotz field to the corresponding far field pattern of the scattered field.

The far field operator F allows a factorization of the form (1) where the operators A and G depend on the specific situation. We will discuss two typical cases and start with the scattering by a sound-soft obstacle D in section "Inverse Acoustic Scattering by a Sound-Soft Obstacle." This is an example of a nonabsorbing medium which is mathematically reflected by the fact that the far field operator is normal – though not self-adjoint as for the corresponding problem in impedance tomography. It was this example for which the factorization method was developed for the first time in [66]. In section "Inverse Electromagnetic Scattering by an Inhomogeneous Medium," we will study the scattering of time-harmonic electromagnetic plane waves by an absorbing medium. In this case the corresponding far field operator fails to be normal.

Each case study will start with a short repetition of the corresponding direct problem. Then the inverse problem will be stated, and a factorization of the form

(1) will be derived. As in impedance tomography, a crucial point is to establish in each case a certain coercivity condition for G. In addition, one needs to prove a range identity which describes the range of A via the known – possibly non-normal – data operator F.

Here and throughout the following sections, $S^2 = \{x \in \mathbb{R}^3 : |x| = 1\}$ denotes the unit sphere in $\mathbb{R}^3$.

Inverse Acoustic Scattering by a Sound-Soft Obstacle

This section is devoted to the analysis of the factorization method for the most simplest case in scattering theory. We consider the scattering of time-harmonic plane waves by an impenetrable obstacle $D \subset \mathbb{R}^3$ which we model by assuming Dirichlet boundary conditions on the boundary ∂D of D. As before, we assume that D is a finite union $D = \cup_{i=1}^m D_i$ of bounded domains D_i such that $\overline{D_i} \cap \overline{D_j} = \emptyset$ for $i \neq j$. Furthermore, we assume that the boundaries ∂D_i are Lipschitz continuous and that the exterior $\mathbb{R}^3 \setminus \overline{D}$ of $\overline{D}$ is connected. Finally, let $k > 0$ be the wave number and

$$u^i(x) = \exp(\mathrm{i}kx \cdot \hat{\theta}), \quad x \in \mathbb{R}^3, \tag{42}$$

be the incident plane wave of direction $\hat{\theta} \in S^2$. The obstacle D gives rise to a scattered field $u^s \in C^2(\mathbb{R}^3 \setminus \overline{D}) \cap C(\mathbb{R}^3 \setminus D)$ which superposes u^i and results in the total field $u = u^i + u^s$ which satisfies the *Helmholtz equation*

$$\Delta u + k^2 u = 0 \quad \text{outside } \overline{D} \tag{43}$$

and the Dirichlet boundary condition

$$u = 0 \quad \text{on } \partial D. \tag{44}$$

The scattered field u^s satisfies the *Sommerfeld radiation condition*

$$\frac{\partial u^s}{\partial r} - \mathrm{i}ku^s = \mathcal{O}(r^{-2}) \quad \text{for } r = |x| \to \infty \tag{45}$$

uniformly with respect to $\hat{x} = x/|x| \in S^2$.

The **direct scattering problem** is to determine the scattered field u^s for a given obstacle $D \subset \mathbb{R}^3$, some $\hat{\theta} \in S^2$ and $k > 0$.

For the treatment of this direct problem, we refer to [36] (see also section "Obstacle Scattering" in chapter ▸ Inverse Scattering). There it is also shown that the scattered field u^s has the asymptotic behavior

$$u^s(x) = \frac{\exp(\mathrm{i}k|x|)}{4\pi|x|} u^\infty(\hat{x}) + \mathcal{O}(|x|^{-2}), \quad |x| \to \infty \tag{46}$$

uniformly with respect to $\hat{x} = x/|x| \in S^2$. The function $u^\infty : S^2 \to \mathbb{C}$ is analytic and is called the *far field pattern* of u^s. It depends on the wave number k, the direction $\hat{\theta} \in S^2$, and the domain D. Since we will keep $k > 0$ fixed, only the dependence on $\hat{\theta}$ is indicated: $u^\infty = u^\infty(\hat{x}; \hat{\theta})$ for $\hat{x}, \hat{\theta} \in S^2$.

In the **inverse scattering problem**, the far field pattern $u^\infty(\hat{x}; \hat{\theta})$ is known for all $\hat{x}, \hat{\theta} \in S^2$ and some fixed $k > 0$, and the domain D has to be determined. We refer again to [36] or chapter ▸ Inverse Scattering for the presentation of the most important properties of this inverse scattering problem. The knowledge of $u^\infty(\hat{x}; \hat{\theta})$ for all $\hat{x}; \hat{\theta} \in S^2$ determines the integral kernel of the *far field operator* F from $L^2(S^2)$ into itself, which is defined by

$$(Fg)(\hat{x}) = \int_{S^2} u^\infty(\hat{x}; \hat{\theta}) g(\hat{\theta}) \mathrm{d}s(\hat{\theta}) \quad \text{for } \hat{x} \in S^2. \tag{47}$$

The far field operator F is compact, normal (i.e., F commutes with its adjoint F^*), and the so-called *scattering operator* $I + \frac{ik}{8\pi^2} F$ is unitary in $L^2(S^2)$.

As in section "Conducting Obstacles," the first step is to derive a factorization of F in the form (1).

The operator A is the *data-to-pattern operator* which maps $f \in H^{1/2}(\partial D)$ to the far field pattern v^∞ of the radiating (i.e., v satisfies the Sommerfeld radiation condition (45)) solution $v \in H^1_{loc}(\mathbb{R}^3 \setminus \overline{D})$ of

$$\Delta v + k^2 v = 0 \text{ in the exterior of } \overline{D}, \; v = f \text{ on } \partial D. \tag{48}$$

Here, $H^1_{loc}(\mathbb{R}^3 \setminus \overline{D})$ is the space of functions v with $v|_{B \setminus \overline{D}} \in H^1(B \setminus \overline{D})$ for all balls $B \subset \mathbb{R}^3$. Existence and uniqueness is assured (see, e.g., [80]; Chap. 9).

Theorem 10. *Define the operator $A : H^{1/2}(\partial D) \to L^2(S^2)$ by $Af = v^\infty$ where v^∞ is the far field pattern of the unique radiating solution $v \in H^1_{loc}(\mathbb{R}^3 \setminus \overline{D})$ of (48). Then A is one-to-one with dense range, and the following factorization holds*

$$F = -AS^* A^*, \tag{49}$$

where $A^* : L^2(S^2) \to H^{-1/2}(\partial D)$ is the dual of A, and $S^* : H^{-1/2}(\partial D) \to H^{1/2}(\partial D)$ is the dual of the single-layer boundary operator $S : H^{-1/2}(\partial D) \to H^{1/2}(\partial D)$ defined by

$$(S\varphi)(x) = \int_{\partial D} \varphi(y) \Phi(x, y) \mathrm{d}s(y), \quad x \in \partial D. \tag{50}$$

Here, Φ denotes the fundamental solution of the Helmholtz equation, i.e.,

$$\Phi(x, y) = \frac{\exp(ik|x - y|)}{4\pi|x - y|}, \qquad x, y \in \mathbb{R}^3, x \neq y, \tag{51}$$

and the explicit definition (50) of this operator makes only sense for smooth functions φ. It has to be extended to functionals $\varphi \in H^{-1/2}(\partial D)$ by a density or duality argument.

Proof. The injectivity of A follows immediately from Rellich's Lemma (see [36] or chapter ▸ Inverse Scattering). The denseness of the range of A can be shown by approximating any $g \in L^2(S^2)$ by a finite sum of spherical harmonics to which the corresponding field can be written down explicitly.

To derive the factorization, define the auxiliary operator $\mathcal{H} : L^2(S^2) \to H^{1/2}(\partial D)$ by

$$(\mathcal{H}g)(x) = \int_{S^2} g(\hat{\theta})\ \exp(\mathrm{i}kx\cdot\hat{\theta})\mathrm{d}s(\hat{\theta}) = \int_{S^2} g(\hat{\theta})u^i(x;\hat{\theta})\mathrm{d}s(\hat{\theta}), \quad x \in \partial D.$$

First we note that $u^\infty(\cdot;\hat{\theta}) = -Au^i(\cdot;\hat{\theta})$ by the definition of A, and thus by the superposition principle, $Fg = -A\mathcal{H}g$ for all $g \in L^2(S^2)$, i.e., $F = -A\mathcal{H}$. We compute the dual $\mathcal{H}^* : H^{-1/2}(\partial D) \to L^2(S^2)$ as

$$(\mathcal{H}^*\varphi)(\hat{x}) = \int_{\partial D} \varphi(y)\exp(-\mathrm{i}k\hat{x}\cdot y)\mathrm{d}s(y), \quad \hat{x} \in S^2.$$

The fundamental solution Φ has the asymptotic behavior

$$\Phi(x,y) = \frac{\exp(\mathrm{i}k|x|)}{4\pi|x|}\exp(-\mathrm{i}k\hat{x}\cdot y) + \mathcal{O}(|x|^{-2}),\ |x| \to \infty \tag{52}$$

uniformly with respect to $\hat{x} \in S^2$ and $y \in \partial D$ and thus has the far field pattern $\Phi^\infty(\hat{x},y) = \exp(-\mathrm{i}k\hat{x}\cdot y)$. Therefore, again by superposition, $\mathcal{H}^*\varphi = AS\varphi$, i.e., $\mathcal{H} = S^*A^*$. Substituting this into $F = -A\mathcal{H}$ yields (49). ∎

Therefore, F allows a factorization in the form (1) with $G = -S^*$. The most important properties of this operator are collected in the following theorem. (For a proof see, e.g., [73, 80].)

Theorem 11. *Assume that k^2 is not a Dirichlet eigenvalue of $-\Delta$ in D. Then the following holds:*

- *S is an isomorphism from the Sobolev space $H^{-1/2}(\partial D)$ onto $H^{1/2}(\partial D)$.*
- *$\mathrm{Im}\langle\varphi,\ S\varphi\rangle < 0$ for all $\varphi \in H^{-1/2}(\partial D)$ with $\varphi \neq 0$. Here, $\langle .,. \rangle$ denotes the duality pairing in $\langle H^{-1/2}(\partial D),\ H^{1/2}(\partial D)\rangle$.*
- *Let S_i be the single-layer boundary operator (50) corresponding to the wave number $k = \mathrm{i}$. The operator S_i is self-adjoint and coercive as an operator from $H^{-1/2}(\partial D)$ onto $H^{1/2}(\partial D)$, i.e., there exists $c_0 > 0$ with*

$$\langle\varphi S_\mathrm{i}\varphi\rangle \geq c_0\,\|\varphi\|^2_{H^{-1/2}(\partial D)} \quad \text{for all } \varphi \in H^{-1/2}(\partial D). \tag{53}$$

- *The difference $S - S_i$ is compact from $H^{-1/2}(\partial D)$ into $H^{-1/2}(\partial D)$.*

From this theorem the following coercivity result can be derived.
Assume that k^2 is not a Dirichlet eigenvalue of $-\Delta$ in D. Then there exists $c_1 > 0$ with

$$|\langle \varphi S\varphi \rangle| \geq c_1 \|\varphi\|^2_{H^{-1/2}(\partial D)} \quad \text{for all } \varphi \in H^{-1/2}(\partial D). \tag{54}$$

This establishes the first step of the factorization method. In the second step, the domain D is characterized by the range of the operator A.

Theorem 12. *For any $z \in \mathbb{R}^3$, define the function $\phi_z \in L^2(S^2)$ by*

$$\phi_z(\hat{x}) = \exp(-ik\hat{x} \cdot z), \ \hat{x} \in S^2. \tag{55}$$

Then z belongs to D, if and only if $\phi_z \in \mathcal{R}(A)$.

Proof. Let first $z \in D$. From (52), we conclude that ϕ_z is the far field pattern of $\Phi(\cdot, z)$; thus, $\phi_z = Af$, where $f = \Phi(\cdot, z)|_{\partial D} \in H^{1/2}(\partial D)$.

Let now $z \notin D$ and assume, on the contrary, that $\phi_z = Af$ for some $f \in H^{1/2}(\partial D)$. Let v be as in the definition of Af. Then $\phi_z = v^\infty$. From Rellich's Lemma and unique continuation, we conclude that $\Phi(\cdot, z)$ and v coincide in $\mathbb{R}^3 \setminus (\overline{D} \cup \{z\})$. By the same arguments as in the proof of Theorem 1, this is a contradiction since v is regular and $\Phi(\cdot, z)$ is singular at z. ■

From the factorization (49), we conclude that $\mathcal{R}(F) \subset \mathcal{R}(A)$, and thus

$$\phi_z \in \mathcal{R}(F) \Longrightarrow z \in D.$$

Therefore, the condition on the left-hand side determines only a subset of D. One can show, cf. [35], that for the case of D, being a ball, the left-hand side is only satisfied for the center of this ball. Nevertheless, the (regularized version) of the test $\phi_z \in \mathcal{R}(F)$ leads to the *Linear Sampling Method*, cf. section "The Linear Sampling Method".

In the third step of the factorization method, the range $\mathcal{R}(A)$ of A has to be expressed by the known data operator F. This is achieved by a second factorization of F based on the spectral decomposition of the normal operator F. From now on we make the assumption that k^2 is not a Dirichlet eigenvalue of $-\Delta$ in D. Then the far field operator is one-to-one as it follows directly from the factorization (49) and part (a) of Theorem 11.

Since F is compact, normal, and one-to-one, there exists a complete set of orthonormal eigenfunctions $\psi_j \in L^2(S^2)$ with corresponding eigenvalues $\lambda_j \in \mathbb{C}$, $j = 1, 2, 3, \ldots$ (see, e.g., [88]). Furthermore, since the operator $I + ik/(8\pi^2)F$ is unitary, the eigenvalues λ_j of F lie on the circle of radius $1/r$ and center i/r

where $r = k/(8\pi^2)$. The spectral theorem for normal operators yields that F has the form

$$F\psi = \sum_{j=1}^{\infty} \lambda_j (\psi, \psi_j)_{L^2(S^2)} \psi_j, \quad \psi \in L^2(S^2). \tag{56}$$

Therefore, F has a second factorization in the form

$$F = (F^*F)^{1/4} G_2 (F^*F)^{1/4}, \tag{57}$$

where the self-adjoint operator $(F^*F)^{1/4} : L^2(S^2) \to L^2(S^2)$ and the signum $G_2 : L^2(S^2) \to L^2(S^2)$ of F are given by

$$(F^*F)^{1/4}\psi = \sum_{j=1}^{\infty} \sqrt{|\lambda_j|}(\psi, \psi_j)_{L^2(S^2)} \psi_j, \quad \psi \in L^2(S^2), \tag{58}$$

$$G_2\psi = \sum_{j=1}^{\infty} \tfrac{\lambda_j}{|\lambda_j|}(\psi, \psi_j)_{L^2(S^2)} \psi_j, \quad \psi \in L^2(S^2). \tag{59}$$

Also this operator G_2 satisfies a coercivity condition of the form (54).

Theorem 13. *Assume that k^2 is not a Dirichlet eigenvalue of $-\Delta$ in D. Then there exists $c_2 > 0$ with*

$$\left|(\psi, G_2\psi)_{L^2(S^2)}\right| \geq c_2 \|\psi\|^2_{L^2(S^2)} \quad \textit{for all } \psi \in L^2(S^2). \tag{60}$$

Proof. It is sufficient to prove (60) for $\psi \in L^2(S^2)$ of the form $\psi = \sum_j c_j \psi_j$ with $\|\psi\|^2_{L^2(S^2)} = \sum_j |c_j|^2 = 1$. With the abbreviation $s_j = \lambda_j/|\lambda_j|$, it is

$$\left|(G_2\psi, \psi)_{L^2(S^2)}\right| = \left| \left(\sum_{j=1}^{\infty} s_j c_j \psi_j, \sum_{j=1}^{\infty} c_j \psi_j \right)_{L^2(S^2)} \right| = \left| \sum_{j=1}^{\infty} s_j |c_j|^2 \right|.$$

The complex number $\sum_{j=1}^{\infty} s_j |c_j|^2$ belongs to the closure of the convex hull $\mathcal{C} = \operatorname{conv}\{s_j : j \in \mathbb{N}\} \subset \mathbb{C}$ of the complex numbers s_j. We conclude that

$$\left|(G_2\psi, \psi)_{L^2(S^2)}\right| \geq \inf\{|z| : z \in \mathcal{C}\}$$

for all $\psi \in L^2(S^2)$ with $\|\psi\|_{L^2(S^2)} = 1$. From the facts that λ_j lie on the circle with center i/r passing through the origin and that λ_j tends to zero as j tends to infinity, we conclude that the only accumulation points of the sequence $\{s_j\}$ can be $+1$ or -1. From the factorization (49) and Theorem 11, it can be shown (see the proof of Theorem 1.23 of [73]) that indeed 1 is the only accumulation point, i.e., $s_j \to 1$ as

j tends to infinity. Therefore, the set $\mathcal{C}$ is contained in the part of the upper half disk which is above the line $l = \{t\hat{s} + (1-t)1 : t \in \mathbb{R}\}$ passing through $\hat{s}$ and 1. Here, $\hat{s}$ is the point in $\{s_j : j \in \mathbb{N}\}$ with the smallest real part. Therefore, the distance of the origin to this convex hull $\mathcal{C}$ is positive, i.e., there exists c_2 with (60).

From Theorem 10 and Eq. (57), the scattering operator F can be written as

$$F = AG_1A^* = (F^*F)^{1/4}G_2(F^*F)^{1/4}, \tag{61}$$

where we have set $G_1 = -S^*$. Both of the operators G_j, $j = 1, 2$ are coercive in the sense of (54) and (60), respectively. By the range identity of Corollary 4, the ranges of A and $(F^*F)^{1/4}$ coincide. The combination of this result and Theorem 12 yields the main result of this section. (To derive the second equivalence of (62), Theorem 25 of Picard has been applied.)

Theorem 14. *Assume that k^2 is not a Dirichlet eigenvalue of $-\Delta$ in D. For any $z \in \mathbb{R}^3$, define again $\phi_z \in L^2(S^2)$ by (55), i.e.,*

$$\phi_z(\hat{x}) := \exp(-\mathrm{i}k\hat{x}\cdot z),\ \hat{x} \in S^2.$$

Then

$$z \in D \iff \phi_z \in \mathcal{R}((F^*F)^{1/4}) \iff \sum_j \frac{\left|(\phi_z, \psi_j)_{L^2(S^2)}\right|^2}{|\lambda_j|} < \infty. \tag{62}$$

Here, $\lambda_j \in \mathbb{C}$ are the eigenvalues of the normal operator F with corresponding normalized eigenfunctions $\psi_j \in L^2(S^2)$.

Formula (62) provides a simple and fast technique to visualize the object D by plotting the inverse of the series on the right-hand side. In practice, this will be a finite sum instead of a series, but the value of the finite sum is much larger for points z outside than for points inside of D. We refer to the original paper [66] for some typical plots.

Remark 2. It is illuminating to compare the presentation in this section with the one for impedance tomography from section "Conducting Obstacles". The relative potential $u - u_{\mathbb{1}}$ considered there corresponds to the scattered wave $u^s = u - u^i$, i.e., the total field minus the incoming field; the incoming field is the potential that is induced by the excitation if the background is homogeneous, whereas the total field is the corresponding solution in the presence of the scatterer.

In both cases, the operator that maps the excitation onto the associated "relative data" can be factorized in three operators: the one that is applied first, i.e., A^*, maps the excitation/the incoming field onto the boundary of the obstacle(s), the operator A that is applied last, maps appropriate boundary data on the obstacle onto the "outgoing field," and its measured data. Accordingly, the operator in the middle encodes the "refraction" at the obstacle(s).

As such, we can view the factorization from impedance tomography as a generalization of Huygens' principle to the diffusion problem (22), although the time causality from scattering theory has no apparent physical analog in stationary diffusion processes.

Inverse Electromagnetic Scattering by an Inhomogeneous Medium

This section is devoted to the analysis of the factorization method for the inverse scattering of electromagnetic time-harmonic plane waves by an inhomogeneous nonmagnetic and conducting medium. Let $k = \omega\sqrt{\varepsilon_0\mu_0} > 0$ be the *wave number* with angular frequency ω, electric permittivity ε_0, and magnetic permeability μ_0 in vacuum. The incident plane wave has the form

$$H^i(x) = p\ \exp\,(\mathrm{i}k\hat{\theta}\cdot x), \quad E^i(x) = -\frac{1}{\mathrm{i}\omega\varepsilon_0}\operatorname{curl} H^i(x), \tag{63}$$

for some polarization vector $p \in \mathbb{C}^3$ and some direction $\hat{\theta} \in S^2$ such that $p \cdot \hat{\theta} = 0$. This pair satisfies the time-harmonic Maxwell system in vacuum, i.e.,

$$\operatorname{curl} E^i - \mathrm{i}\omega\mu_0 H^i = 0 \quad \text{in } \mathbb{R}^3, \tag{64}$$

$$\operatorname{curl} H^i + \mathrm{i}\omega\varepsilon_0 E^i = 0 \quad \text{in } \mathbb{R}^3. \tag{65}$$

This incident wave is scattered by a medium with space-dependent electric permittivity $\varepsilon = \varepsilon(x)$ and conductivity $\sigma = \sigma(x)$. We assume that the magnetic permeability μ is constant and equal to the permeability μ_0 of vacuum. Furthermore, we assume that $\varepsilon \equiv \varepsilon_0$ and $\sigma \equiv 0$ outside of some bounded domain. The total fields are superpositions of the incident and scattered fields, i.e., $E = E^i + E^s$ and $H = H^i + H^s$ and satisfy the Maxwell system

$$\operatorname{curl} E - \mathrm{i}\omega\mu_0 H = 0 \quad \text{in } \mathbb{R}^3, \tag{66}$$

$$\operatorname{curl} H + \mathrm{i}\omega\varepsilon E = \sigma E \quad \text{in } \mathbb{R}^3. \tag{67}$$

Also, the tangential components of E and H are continuous on interfaces where σ or ε are discontinuous. Finally, the scattered fields have to be radiating, i.e., satisfy the *Silver-Müller radiation condition*

$$\sqrt{\mu_0}H^s(x) \times \hat{x} - \sqrt{\varepsilon_0}E^s(x) = \mathcal{O}\left(\frac{1}{|x|^2}\right) \quad \text{as } |x| \to \infty \tag{68}$$

uniformly w.r.t. $\hat{x} = x/|x| \in S^2$. The complex-valued *relative electric permittivity* ε_r is defined by

$$\varepsilon_r(x) = \frac{\varepsilon(x)}{\varepsilon_0} + \mathrm{i}\frac{\sigma(x)}{\omega\varepsilon_0}. \tag{69}$$

Note that $\varepsilon_r \equiv 1$ outside of some bounded domain. Equation (67) can then be written in the form

$$\operatorname{curl} H + \mathrm{i}\omega\varepsilon_0\varepsilon_r E = 0 \quad \text{in } \mathbb{R}^3. \tag{70}$$

It is preferable to work with the magnetic field H only. This is motivated by the fact that the magnetic field is divergence free as seen from (66) and the fact that div curl $= 0$. In general, this is not the case for the electric field E. Eliminating the electric field E from the system (66) and (70) leads to

$$\operatorname{curl}\left[\frac{1}{\varepsilon_r}\operatorname{curl} H\right] - k^2 H = 0 \quad \text{in } \mathbb{R}^3. \tag{71}$$

The incident field H^i satisfies

$$\operatorname{curl}^2 H^i - k^2 H^i = 0 \quad \text{in } \mathbb{R}^3. \tag{72}$$

Subtracting both equations yields

$$\operatorname{curl}\left[\frac{1}{\varepsilon_r}\operatorname{curl} H^s\right] - k^2 H^s = \operatorname{curl}\left[q \operatorname{curl} H^i\right] \quad \text{in } \mathbb{R}^3, \tag{73}$$

where the contrast q is defined by $q = 1 - 1/\varepsilon_r$. The Silver-Müller radiation condition turns into

$$\operatorname{curl} H^s(x) \times \hat{x} - \mathrm{i}kH^s(x) = \mathcal{O}\left(\frac{1}{|x|^2}\right), \quad |x| \to \infty. \tag{74}$$

The continuity of the tangential components of E and H translates into analogous requirements for H^s and curl H^s.

It will be necessary to allow more general source terms on the right-hand side of (73). In particular, we will consider the problem to determine a radiating solution $v \in H_{loc}(\operatorname{curl}, \mathbb{R}^3)$ of

$$\operatorname{curl}\left[\frac{1}{\varepsilon_r}\operatorname{curl} v\right] - k^2 v = \operatorname{curl} f \quad \text{in} \quad \mathbb{R}^3 \tag{75}$$

for given $f \in L^2(\mathbb{R}^3)^3$ with compact support. (For any open set D the space $L^2(D)^3$ denotes the space of vector functions $v : D \to \mathbb{C}^3$ such that all components are in $L^2(D)$.) The solutions v of (75) as well as of (71) and (73) have to be understood in the variational sense, i.e., $v \in H_{loc}(\operatorname{curl}, \mathbb{R}^3)$ satisfies

$$\int_{\mathbb{R}^3} \left[\frac{1}{\varepsilon_r} \operatorname{curl} v \cdot \operatorname{curl} \psi - k^2 v \cdot \psi \right] \mathrm{d}x = \int_{\mathbb{R}^3} f \cdot \operatorname{curl} \psi \, \mathrm{d}x \tag{76}$$

for all $\psi \in H(\operatorname{curl}, \mathbb{R}^3)$ with compact support. For any domain Ω, the Sobolev space $H(\operatorname{curl}, \Omega)$ is the space of all vector fields $v \in L^2(\Omega)^3$ such that also $\operatorname{curl} v \in L^2(\Omega)^3$. Furthermore, $H_{loc}(\operatorname{curl}, \mathbb{R}^3) = \{v : v|_B \in H(\operatorname{curl}, B)$ for all balls $B \subset \mathbb{R}^3\}$.

Outside of the supports of $\varepsilon_r - 1$ and f, the solution satisfies $\operatorname{curl}^2 v - k^2 v = 0$. Taking the divergence of this equation and using the identities $\operatorname{div} \operatorname{curl} = 0$ and $\operatorname{curl}^2 = -\Delta + \operatorname{grad} \operatorname{div}$, this system is equivalent to the pair of equations

$$\Delta v + k^2 v = 0 \quad \text{and} \quad \operatorname{div} v = 0.$$

Classical interior regularity results (cf. [80] combined with [36]) yield that v is analytic outside of the supports of $\varepsilon_r - 1$ and f. In particular, the radiation condition (74) is well defined.

There are several ways to show the Fredholm property of Eq. (75). We refer to [81] for the treatment by a variational equation with nonlocal boundary conditions or to [73] for a treatment by an integrodifferential equation of Lippmann-Schwinger type.

The question of uniqueness of radiating solutions to (75) is closely related to the validity of the unique continuation principle. It is known to hold for piecewise Hölder-continuously differentiable functions ε_r (see [81]).

As in the case of the Helmholtz equation, every radiating vector field v satisfying $\operatorname{curl}^2 v - k^2 v = 0$ outside of some ball has the asymptotic behavior

$$v(x) = \frac{\exp(\mathrm{i}k|x|)}{4\pi|x|} v^\infty(\hat{x}) + \mathcal{O}(|x|^{-2}), \quad |x| \to \infty,$$

uniformly with respect to $\hat{x} = x/|x| \in S^2$ (see again [36]). The vector field v^∞ is uniquely determined and again called the *far field pattern* of v. It is a tangential vector field, i.e., $v^\infty \in L_t^2(S^2)$ where

$$L_t^2(S^2) = \{w \in L^2(S^2)^3 : w(\hat{x}) \cdot \hat{x} = 0, \hat{x} \in S^2\}.$$

The **inverse problem** is to determine the shape D of the support of the contrast q from the far field pattern $H^\infty(\hat{x}; \hat{\theta}, p)$ for all $\hat{x}, \hat{\theta} \in S^2$ and $p \in \mathbb{C}^3$ with $p \cdot \hat{\theta} = 0$. Because of the linear dependence of H^∞ on p, it is sufficient to know H^∞ only for a basis of two vectors for p. As in impedance tomography, the task of determining only D is rather modest since it is well known that one can even reconstruct q uniquely from this set of data (see [38]). However, the proof of uniqueness is nonconstructive, while the factorization method will provide an explicit characterization of the characteristic function of D which can, e.g., be used for numerical purposes. Also, the factorization method can – with only minor

modifications – be carried over for anisotropic media (as in section "Conducting Obstacles") where it is well known that ε_r can only be determined up to a smooth change of coordinates.

For the remaining part of this section, we make the following assumption:

Assumption 12. *Let $D \subset \mathbb{R}^3$ be a finite union $D = \cup_{i=1}^m D_i$ of bounded domains D_i such that $\overline{D_i} \cap \overline{D_j} = \emptyset$ for $i \neq j$. Furthermore, we assume that the boundaries ∂D_i are Lipschitz continuous and the exterior $\mathbb{R}^3 \setminus \overline{D}$ of $\overline{D}$ is connected. Let $\varepsilon_r \in L^\infty(D)$ be such that the values $\varepsilon_r(x)$ vary in a compact subset of the half disk $\{z \in \mathbb{C} : (\operatorname{Re} z - 1/2)^2 + (\operatorname{Im} z)^2 < 1/4, \operatorname{Im} z \geq 0\}$, and that for every $f \in L^2(\mathbb{R}^3)^3$ with compact support there exists a unique radiating solution of (75).*

We extend ε_r by one outside of D and define the contrast by $q = 1 - 1/\varepsilon_r$; thus, $\operatorname{Im} q \geq 0$ and $\operatorname{Re} q \leq -\gamma|q|$ on D for some $\gamma > 0$.

Condition (3) is, e.g., satisfied for Hölder-continuously differentiable parameters ε and σ (see [81]).

The *far field operator* $F : L_t^2(S^2) \to L_t^2(S^2)$ is defined as

$$(Fp)(\hat{x}) := \int_{S^2} H^\infty(\hat{x}; \theta, p(\theta)) \mathrm{d}s(\theta), \quad \hat{x} \in S^2. \tag{77}$$

F is a linear operator since H^∞ depends linearly on the polarization p.

The first step in the factorization method is to derive a factorization of F in the form $F = AT^*A^*$ where the operators $A : L^2(D)^3 \to L_t^2(S^2)$ and $T : L^2(D)^3 \to L^2(D)^3$ are defined as follows.

The *data-to-pattern operator* $A : L^2(D)^3 \to L_t^2(S^2)$ is defined by $Af := v^\infty$, where v^∞ denotes the far field pattern corresponding to the radiating (variational) solution $v \in H_{loc}(\operatorname{curl}, \mathbb{R}^3)$ of

$$\operatorname{curl}\left[\frac{1}{\varepsilon_r}\operatorname{curl} v\right] - k^2 v = \operatorname{curl}\left[\frac{q}{\sqrt{|q|}} f\right] \quad \text{in } \mathbb{R}^3. \tag{78}$$

Again, the contrast is given by $q = 1 - 1/\varepsilon_r$. We note that the solution exists by part (3) of Assumption 12.

The operator $T : L^2(D)^3 \to L^2(D)^3$ is defined by $Tf = (\operatorname{sign} \overline{q}) f - \sqrt{|q|} \operatorname{curl} w|_D$, where $w \in H_{loc}(\operatorname{curl}, \mathbb{R}^3)$ is the radiating solution of

$$\operatorname{curl}^2 w - k^2 w = \operatorname{curl}\left[\sqrt{|q|} f\right] \quad \text{in } \mathbb{R}^3. \tag{79}$$

The solution exists and is unique (see, e.g., [73]).

Theorem 15. *Let Assumption 12 hold. Then F from (77) can be factorized as*

$$F = AT^*A^*, \tag{80}$$

where $A^* : L_t^2(S^2) \to L^2(D)^3$ *and* $T^* : L^2(D)^3 \to L^2(D)^3$ *denote the adjoints of* A *and* T*, respectively. Furthermore,* A^* *is injective.*

For a proof of this and the following result, we refer to [73].

Remark 3. The solution w of (79) can be expressed in the form (see [73])

$$w(x) = \operatorname{curl} \int_D \sqrt{|q(y)|}\, f(y) \Phi(x, y) \mathrm{d}y, \quad x \in \mathbb{R}^3,$$

which yields an explicit expression of T.

The following theorem corresponds to Theorem 11 and collects properties of the operator T needed for the analysis of the factorization method.

Theorem 16. *Let the conditions of Assumption 12 hold, and let* $T : L^2(D)^3 \to L^2(D)^3$ *be defined above. Then the following holds:*

(a) The imaginary part $\operatorname{Im} T = \frac{1}{2\mathrm{i}}(T - T^*)$ *is non-positive, i.e.,*

$$\operatorname{Im}(Tf, f)_{L^2(D)^3} \leq 0 \quad \text{for all } f \in L^2(D)^3.$$

(b) Define the operator T_0 *in the same way as* T *but for* $k = \mathrm{i}$*. Then* $-\operatorname{Re} T_0$ *is coercive, and* $T - T_0$ *is compact in* $L^2(D)^3$*.*

(c) T *is an isomorphism from* $L^2(D)^3$ *onto itself.*

As in section "Inverse Acoustic Scattering by a Sound-Soft Obstacle" we first characterize the domain D by the range $\mathcal{R}(A)$ of A. The proof of the following result can again be found in [73].

Theorem 17. *Let the conditions of Assumption 12 hold. For any* $z \in \mathbb{R}^3$ *and fixed* $p \in \mathbb{C}^3$*, we define* $\phi_z \in L_t^2(S^2)$ *as the far field pattern of the electric dipole at* z *with moment* p*, i.e.,*

$$\phi_z(\hat{x}) = -\mathrm{i}k(\hat{x} \times p) \exp(-\mathrm{i}k\hat{x} \cdot z), \quad \hat{x} \in S^2. \tag{81}$$

Then z *belongs to* D*, if and only if* $\phi_z \in \mathcal{R}(A)$*.*

In contrast to the data operators $\Lambda_0 - \Lambda_{\mathbb{1}}$ or $\Lambda - \Lambda_{\mathbb{1}}$ of Sect. 2 or the far field operator F of section "Inverse Acoustic Scattering by a Sound-Soft Obstacle", the far field operator for absorbing media – as in the present case – fails to be normal or even self-adjoint. Therefore, the approaches of the previous sections – i.e., the application of the range identities of Corollaries 4 and 5 – are not applicable. However, application of Theorem 27 to the far field operator F from $L_t^2(S^2)$ into itself and the operator $G = T^* : L^2(D)^3 \to L^2(D)^3$ yields the characterization of D via an auxiliary operator

$$F_\# = |\operatorname{Re} F| + \operatorname{Im} F, \tag{82}$$

cf. (109), which is easily obtained from the given far field data.

Theorem 18. *Let the conditions of Assumption 12 hold. For any $z \in \mathbb{R}^3$, define again $\phi_z \in L_t^2(S^2)$ by (81). Then, with $F_\#$ of (82), there holds*

$$z \in D \iff \phi_z \in \mathcal{R}\left(F_\#^{1/2}\right) \iff \sum_j \frac{\left|(\phi_z, \psi_j)_{L^2(S^2)}\right|^2}{|\lambda_j|} < \infty. \tag{83}$$

Here, $\lambda_j \in \mathbb{C}$ are the eigenvalues of the self-adjoint and positive compact operator $F_\#$ with corresponding normalized eigenfunctions $\psi_j \in L_t^2(S^2)$.

Historical Remarks and Open Questions

Historically, the factorization method originated from the linear sampling method which will be explained in section "The Linear Sampling Method" (see also section "The Linear Sampling Method" in chapter ▸ Inverse Scattering). The linear sampling method studies the *far field equation* $Fg = \phi_z$ in contrast to the factorization method which characterizes the domain D by *exactly* those points z for which the modified far field equation $F_\#^{1/2} g = \phi_z$ is solvable where $F_\# = (F^* F)^{1/2}$ in the case of section "Inverse Acoustic Scattering by a Sound-Soft Obstacle" and $F_\# = |\operatorname{Re} F| + \operatorname{Im} F$ in the case of section "Inverse Electromagnetic Scattering by an Inhomogeneous Medium." It is easily seen that the points for which the far field equation $Fg = \phi_z$ is solvable determines only a subset of D – which can consist of a single point only, as the example of a ball shows.

The implementation of the factorization method is as simple and universal as of the linear sampling method. Only the far field operator F – i.e., in practice a finite-dimensional approximation – has to be known. No other a priori information on the unknown domain D such as the number of components or the kind of boundary condition has to be known in advance. The mathematical justification, however, has to be proven for every single situation. Since their first presentations, the factorization method has been justified for several problems in inverse acoustic and electromagnetic scattering theory such as the scattering by inhomogeneous media [67, 69, 72, 73], scattering by periodic structures [11, 12], and scattering by obstacles under different kinds of boundary conditions [49, 73]. The factorization method can also be adapted for scattering problems for a crack [74] with certain modifications; we refer to the remarks concerning the crack problem in section "Other Generalizations." The factorization method for elastic scattering problems and wave guides is studied in [9] and [30], respectively.

In many situations near field measurements on some surface Γ for point sources on the same surface Γ as incident fields rather than far field measurements for plane waves as incident fields are available. The corresponding "near field operator" M :

$L^2(\Gamma) \to L^2(\Gamma)$ allows a factorization in the form $M = BGB'$ where B' is the adjoint with respect to the bilinear form $\int_\Gamma uv\mathrm{d}s$ rather than the (sesquilinear) inner product $\int_\Gamma u\bar{v}\mathrm{d}s$. The validity of the range identity for these kinds of factorizations is not known so far and is one of the open problems in this field. For certain situations (see [73]), the corresponding far field operator F can be computed from M, and the factorization method can then be applied to F.

Also the cases where the background medium is more complicated than the free space can be treated (see [48,73] for scattering problems in a half space and [71] for scattering problems in layered media).

The justification of the factorization method for arbitrary elliptic boundary value problems or even more general problems is treated in [44,70,82].

4 Related Sampling Methods

This section is devoted to some alternate examples of sampling methods which were developed during the last decade: the *linear sampling method*, first introduced by Colton and Kirsch in [35], the closely related *MUSIC*, the *singular sources method* by Potthast (see [85]), and Ikehata's *probe method* (see [62]). However, it is not the aim of this section to report on all sampling methods. In particular, we do not discuss the *enclosure method* or the *no-response test* but refer to the monograph [86] and the survey article [87].

The Linear Sampling Method

Here we reconsider the inverse scattering problem for time-harmonic plane acoustic waves of section "Inverse Acoustic Scattering by a Sound-Soft Obstacle," i.e., the problem to determine the shape of an acoustically soft obstacle D from the knowledge of the far field pattern $u^\infty(\hat{x};\hat{\theta})$ for all $\hat{x},\hat{\theta} \in S^2$. We refer to (42)–(47) for the mathematical model and the definition of the far field operator F from $L^2(S^2)$ into itself.

The factorization method for inverse scattering problems studies solvability of the equation $F_\#^{1/2} g = \phi_z$ in $L^2(S^2)$ where $F_\# = (F^*F)^{1/2}$ in the case where F is normal (as, e.g., in section "Inverse Acoustic Scattering by a Sound-Soft Obstacle") and $F_\# = |\operatorname{Re} F| + \operatorname{Im} F$ in the general case with absorption (see Theorems 14 and 18, respectively). In contrast to this equation, the *linear sampling method* considers the *far field equation*

$$Fg = \phi_z \quad \text{in } L^2(S^2). \tag{84}$$

We mention again that in general no solution of this equation exists. However, one can compute "approximate solutions" $g = g_{z,\varepsilon}$ of (84) such that $\| g \|_{L^2(S^2)}$ behaves differently for z being inside or outside of D. We refer to chapter ▸ Inverse Scattering, Theorem 5.3, for a more precise formulation of this behavior.

The drawback of this result – and all the other attempts to justify the linear sampling method rigorously – is that there is no guarantee that the solution of a regularized version of (84), e.g., by Tikhonov regularization, will actually pick the density $g = g_{z,\varepsilon}$ with the properties of the aforementioned "approximate solution." We refer to [53] for a discussion of this fact. However, numerically the method has proven to be very effective for a large class of inverse scattering problems (see, e.g., [26] for the scattering by cracks, [27] for inverse scattering problems for anisotropic media, [19] for wave guide scattering problems, [33, 34, 51] for electromagnetic scattering problems, and [29, 31, 40] for elastic scattering problems). Modifications of the linear sampling method and combinations with other methods can be found in [8, 20, 79].

For the cases in which the factorization method in the form $(F^*F)^{1/4}g = \phi_z$ is applicable, a complete characterization of the unknown obstacle D by a modification of the linear sampling method can be derived by replacing the indicator value $\|g\|_{L^2(S^2)}$ by $(g, \phi_z)_{L^2(S^2)}$. This is summarized in the following theorem (see [10, 13] and, for the following presentation, [73]).

Theorem 19. *Let $u^\infty = u^\infty(\hat{x}; \hat{\theta})$ be the far field pattern corresponding to the scattering problem (42)–(45) with associated far field operator F, and assume that k^2 is not a Dirichlet eigenvalue of $-\Delta$ in D. Furthermore, for every $z \in D$, let $g_z \in L^2(S^2)$ denote the solution of $(F^*F)^{1/4} g_z = \phi_z$, i.e., the solution obtained by the factorization method, and for every $z \in \mathbb{R}^3$ and $\varepsilon > 0$, let $g = g_{z,\varepsilon} \in L^2(S^2)$ be the Tikhonov approximation of (84), i.e., the unique solution of*

$$(\varepsilon I + F^*F)g = F^*\phi_z \tag{85}$$

which is computed by the linear sampling method (if Tikhonov's regularization technique is chosen). Here, $\phi_z \in L^2(S^2)$ is defined in (55). Furthermore, let $v_{g_{z,\varepsilon}}(z) = (g_{z,\varepsilon}, \phi_z)_{L^2(S^2)} = \int_{S^2} g_{z,\varepsilon}(\hat{\theta}) \exp(\mathrm{i}k\hat{\theta} \cdot z) \mathrm{d}s(\hat{\theta})$ denote the corresponding Herglotz wave function evaluated at z.

(a) *For every $z \in D$, the limit $\lim_{\varepsilon \to 0} v_{g_{z,\varepsilon}}(z)$ exists. Furthermore, there exists $c > 0$, depending on F only, such that for all $z \in D$ the following estimates hold:*

$$c \, \|g_z\|^2_{L^2(S^2)} \leq \lim_{\varepsilon \to 0} \left| v_{g_{z,\varepsilon}}(z) \right| \leq \|g_z\|^2_{L^2(S^2)} \, . \tag{86}$$

(b) *For $z \notin D$, the absolute values $|v_{g_{z,\varepsilon}}(z)|$ tend to infinity as ε tends to zero.*

Proof. Using an orthonormal system $\{\psi_j : j \in \mathbb{N}\}$ of eigenfunctions ψ_j corresponding to eigenvalues $\lambda_j \in \mathbb{C}$ of F, one computes the Tikhonov approximation $g_{z,\varepsilon}$ from (85) as

$$g_{z,\varepsilon} = \sum_{j=1}^{\infty} \frac{\overline{\lambda_j}}{|\lambda_j|^2 + \varepsilon} (\phi_z, \psi_j)_{L^2(S^2)} \psi_j .$$

From $v_g(z) = (g, \phi_z)_{L^2(S^2)}$ for any $g \in L^2(S^2)$, we conclude that

$$v_{g_{z,\varepsilon}}(z) = \sum_{j=1}^{\infty} \frac{\overline{\lambda_j}}{|\lambda_j|^2 + \varepsilon} \left| (\phi_z, \psi_j)_{L^2(S^2)} \right|^2 . \tag{87}$$

(a) Let now $z \in D$. Then $(F^*F)^{1/4} g_z = \phi_z$ is solvable in $L^2(S^2)$ by Theorem 14, and thus $(\phi_z, \psi_j)_{L^2(S^2)} = ((F^*F)^{1/4} g_z, \psi_j)_{L^2(S^2)} = (g_z, (F^*F)^{1/4} \psi_j)_{L^2(S^2)} = \sqrt{|\lambda_j|} (g_z, \psi_j)_{L^2(S^2)}$.

Therefore, we can express $v_{g_{z,\varepsilon}}(z)$ as

$$v_{g_{z,\varepsilon}}(z) = \sum_{j=1}^{\infty} \frac{\overline{\lambda_j} |\lambda_j|}{|\lambda_j|^2 + \varepsilon} \left| (g_z, \psi_j)_{L^2(S^2)} \right|^2 = \|g_z\|^2_{L^2(S^2)} \sum_{j=1}^{\infty} \rho_j \frac{\overline{\lambda_j} |\lambda_j|}{|\lambda_j|^2 + \varepsilon}, \tag{88}$$

where $\rho_j = \left| (g_z, \psi_j)_{L^2(S^2)} \right|^2 / \|g_z\|^2_{L^2(S^2)}$ is nonnegative with $\Sigma_j \rho_j = 1$. An elementary argument (theorem of dominated convergence) yields convergence

$$\sum_{j=1}^{\infty} \rho_j \frac{\overline{\lambda_j} |\lambda_j|}{|\lambda_j|^2 + \varepsilon} \longrightarrow \sum_{j=1}^{\infty} \rho_j \frac{\overline{\lambda_j}}{|\lambda_j|} = \sum_{j=1}^{\infty} \rho_j \overline{s_j}$$

as ε tends to zero where again $s_j = \lambda_j / |\lambda_j|$. The properties of ρ_j imply that the limit belongs to the closure $\mathcal{C}$ of the convex hull of the complex numbers $\{s_j : j \in \mathbb{N}\}$. The same argument as in the proof of Theorem 13 yields that $\mathcal{C}$ has a positive distance c from the origin, i.e., $\left| \sum_{j=1}^{\infty} \rho_j \overline{s_j} \right| \geq c$ which proves the lower bound. The upper estimate is seen directly from (88).

(b) Let now $z \notin D$, and assume on the contrary that there exists a sequence $\{\varepsilon_n\}$ which tends to zero and such that $|v_n(z)|$ is bounded. Here we have set $v_n = v_{g_{z,\varepsilon_n}}$ for abbreviation. Since s_j converges to 1, there exists $j_0 \in \mathbb{N}$ with $\operatorname{Re} \lambda_j > 0$ for $j \geq j_0$. From (87) for $\varepsilon = \varepsilon_n$, we get

$$v_n(z) = \sum_{j=1}^{j_0 - 1} \frac{\overline{\lambda_j}}{|\lambda_j|^2 + \varepsilon_n} \left| (\phi_z, \psi_j)_{L^2(S^2)} \right|^2 + \sum_{j=j_0}^{\infty} \frac{\overline{\lambda_j}}{|\lambda_j|^2 + \varepsilon_n} \left| (\phi_z, \psi_j)_{L^2(S^2)} \right|^2 .$$

Since the finite sum is certainly bounded for $n \in \mathbb{N}$, there exists $c_1 > 0$ such that

$$\left| \sum_{j=j_0}^{\infty} \frac{\lambda_j}{|\lambda_j|^2 + \varepsilon_n} \left| (\phi_z, \psi_j)_{L^2(S^2)} \right|^2 \right| \leq c_1 \quad \text{for all } n \in \mathbb{N}.$$

Observing that for any complex number $w \in \mathbb{C}$ with $\operatorname{Re} w \geq 0$ and $\operatorname{Im} w \geq 0$ we have that $\operatorname{Re} w + \operatorname{Im} w \geq |w|$, we conclude (note that also $\operatorname{Im} \lambda_j > 0$)

$$2c_1 \leq 2 \left| \sum_{j=j_0}^{\infty} \frac{\lambda_j}{|\lambda_j|^2 + \varepsilon_n} \left| (\phi_z, \psi_j)_{L^2(S^2)} \right|^2 \right| \geq \sum_{j=j_0}^{\infty} \frac{\operatorname{Re} \lambda_j + \operatorname{Im} \lambda_j}{|\lambda_j|^2 + \varepsilon_n} \left| (\phi_z, \psi_j)_{L^2(S^2)} \right|^2$$

$$\geq \sum_{j=j_0}^{\infty} \frac{|\lambda_j|}{|\lambda_j|^2 + \varepsilon_n} \left| (\phi_z, \psi_j)_{L^2(S^2)} \right|^2 \geq \sum_{j=j_0}^{J} \frac{|\lambda_j|}{|\lambda_j|^2 + \varepsilon_n} \left| (\phi_z, \psi_j)_{L^2(S^2)} \right|^2$$

for all $n \in \mathbb{N}$ and all $J \geq j_0$. Letting n tend to infinity yields boundedness of the finite sum uniformly w.r.t. J and thus convergence of the series $\sum_{j=j_0}^{\infty} \frac{1}{|\lambda_j|} \left| (\phi_z, \psi_j)_{L^2(S^2)} \right|^2$. From (62) therefore follows that $z \in D$, which is the desired contradiction. ■

Obviously, this kind of modification of the original linear sampling method can be done for all inverse scattering problems for which Theorem 14 holds. This includes scattering by acoustically hard obstacles or inhomogeneous nonabsorbing media or, with appropriate modifications, scattering by open arcs.

MUSIC

The linear sampling method investigates "to what extent" the far field equation

$$Fg = \phi_z$$

is solvable for a number of sampling points z within some region of interest. As we have mentioned before, this equation has a solution in very rare cases only and usually not for every $z \in D$.

However, if the obstacle is very small, then it turns out that the far field operator almost degenerates to a finite rank operator, in which case the "numerical range" of F and $(F^*F)^{1/4}$ would be the same finite-dimensional subspace, where the latter is known to contain ϕ_z for every $z \in D$ – under appropriate assumptions on the particular problem setting (see Sects. 2 and 3).

To investigate this observation in more detail, we embed the real scene in a parameterized family of problems, where the parameter $\delta > 0$ reflects the scale of the problem. Assume that the scatterer $D = \cup_{i=1}^{m} D_i$ consists of m obstacles given as

$$D_i = z_i + \delta U_i \quad i = 1, \ldots, m, \tag{89}$$

where each domain U_i contains the origin and has Lipschitz continuous boundary and the closure of U_i has a connected complement. We shall call z_i the *location*

of D_i and U_i its *shape*. We focus our presentation on an inhomogeneous medium setting for acoustic scattering, i.e., the Helmholtz equation, to provide analogies to both settings from Sect. 3. Let ρ_0 and c_0 be the density and the speed of sound in vacuum, $k = \omega/c_0$ be the associated wave number with frequency ω, and $u^i(x) = \exp(\mathrm{i}kx\cdot\hat{\theta})$ be an incoming plane wave. Then, if we assume that the density ρ_i and the sound of speed c_i in each object D_i are real and constant, then the total field $u_\delta = u^i + u^s_\delta$ solves the Helmholtz equation (see, e.g., [36])

$$\operatorname{div}\left(\frac{1}{\rho}\operatorname{grad} u_\delta\right) + \omega^2\eta u_\delta = 0 \quad \text{in } \mathbb{R}^3, \tag{90}$$

with the radiation condition

$$\frac{\partial u^s_\delta}{\partial r} - \mathrm{i}k u^s_\delta = \mathcal{O}(r^{-2}) \quad \text{for } r = |x| \to \infty, \tag{91}$$

uniformly with respect to $\hat{x} = x/|x|$, and the parameter η equals $\eta_0 = 1/\rho_0$ in $\mathbb{R}^3\backslash\overline{D}$, and $\eta_i = c_0^2/(c_i^2\rho_i)$ in D_i, $i = 1,\dots,m$, respectively. We mention that for constant $\eta = 1/\rho_0$, it has been shown in [72] that the standard factorization method (with $F_\# = (F^*F)^{1/2}$) applies for this setting with fixed scaling parameter δ. We know of no result, however, where the factorization method is used to reconstruct the supports of $\rho - \rho_0$ and $\eta - \eta_0$ in this setting simultaneously, although there are partial results for a similar problem (in a bounded domain and with a different sign of η) arising in optical tomography, cf. [43, 59].

The idea to approach this problem is based on an asymptotic expansion of the far field u^∞_δ of the scattered wave with respect to the parameter δ in (89). We quote the following result from [4].

Theorem 20. *The far field of the scattering problem (90) and (91) for the scatterers given in (89) satisfies*

$$u^\infty_\delta(\hat{x};\hat{\theta}) = \delta^3 k^2 \sum_{i=1}^{m}\left(\left(\frac{\rho_i}{\rho_0} - 1\right)\hat{x}\cdot M_i\hat{\theta} - \left(\frac{\eta_i}{\eta_0} - 1\right)|U_i|\right)\exp(\mathrm{i}k(\hat{\theta}-\hat{x})\cdot z_i) + o(\delta^3), \tag{92}$$

and the associated far field operator can be rewritten as

$$F = \delta^3\hat{F} + o(\delta^3) \tag{93}$$

in the norm of $\mathcal{L}(L^2(S^2))$, where the rank of the operator $\hat{F}$ is at most $4m$. Here, $|U_i|$ is the Lebesgue measure of U_i, and $M_i \in \mathbb{R}^{3\times 3}$ are symmetric positive definite matrices that depend on the shape U_i, the so-called polarization tensors.

As is obvious, the scattered field and its far field vanish as $\delta \to 0$. The corresponding rate δ^3 reflects the space dimension; in $\mathbb{R}^2$, the corresponding field decays like δ^2 as $\delta \to 0$.

The importance of Theorem 20 stems from the fact that the leading order approximation $\hat{F}$ of the far field operator F has finite rank, whereas F has infinite dimensional range. The rank of $\hat{F}$ is $4m$, unless some of the scatterers have the same material parameters as the background vacuum. Note that the dominating term of u_δ^∞ consists of two parts: the first contribution stems from the change in the density ρ and corresponds to the far field of a dipole (point source) in z_i; likewise, the second term corresponds to the far field of a monopole in z_i, and this is the result of a change in the parameter η.

It is easy to deduce from Theorem 20 that we can factorize $\hat{F}$ quite naturally in three factors.

Theorem 21. *The operator $\hat{F} : L^2(S^2) \to L^2(S^2)$ admits a factorization of the form*

$$\hat{F} = -BMB', \tag{94}$$

where $B : \mathbb{C}^{4m} \to L^2(S^2)$ maps a vector $[p_1, \ldots, p_m,\ a_1, \ldots, a_m]^T \in \mathbb{C}^{4m}$ with $p_i \in \mathbb{C}^3$ and $a_i \in \mathbb{C}$, $i = 1, \ldots, m$, to the far field of

$$u(x) = \sum_{i=1}^{m} (p_i \cdot \operatorname{grad}_z \Phi(x, z_i) + a_i \Phi(x, z_i)),$$

where Φ is as in (51), $M \in \mathbb{R}^{4m \times 4m}$ is a real block diagonal matrix with m blocks of size 3×3 and m single elements on its diagonal, and M is nonsingular, if and only if $\rho_i \neq \rho_0$ and $\eta_i \neq \eta_0$ for all $i = 1, \ldots, m$. The operator B' is the dual operator of B with respect to the bilinear forms of $\mathbb{C}^{4m}$ and $L^2(S^2)$, i.e., $B'g$ consists of the gradients and point values of the Herglotz wave function

$$v_g(x) = \int_{S^2} g(\hat{\theta}) \exp(\mathrm{i}kx \cdot \hat{\theta}) \mathrm{d}s(\hat{\theta}), \qquad x \in \mathbb{R}^3,$$

evaluated at the points z_i, $i = 1, \ldots, m$.

As M in (94) is invertible, the range of $\hat{F}$ and the range of B coincide, and it consists of the far fields of the monopoles and all possible dipoles emanating from the locations z_i of D_i, $i = 1, \ldots, m$. Using the unique continuation principle, we can thus conclude the following result.

Corollary 2. *If each scatterer has a different parameter η than the background medium, then a point $z \in \mathbb{R}^3$ is the location z_i of one of the scatterers, if and only if ϕ_z of (55) belongs to the range of $\hat{F}$.*

When δ is small, it follows from (93) that numerically the range of F and the range of $\hat{F}$ are the same, essentially. By this we mean that the dominating $4m$ singular values of F are small perturbations of the nonzero singular values of $\hat{F}$, and the corresponding singular subspaces are also close to each other. Moreover, we expect to see a sharp gap between the $4m$th and the $4m+1$st singular value of F. We can search for this gap to determine the number m of the scatterers and then determine the angle between the test function ϕ_z and the $4m$-dimensional dominating singular subspace of F. When z is close to the location of one of the scatterers, then this angle will be small, otherwise this angle will be larger. This way images can be produced that enable one to visualize the approximate locations of the scatterers, but not their shape.

This approach applies for all problem settings that have been discussed in Sects. 2 and 3 and many more. In impedance tomography, for example, the corresponding asymptotic expansion of the boundary potential has the form

$$u_\delta(x) - u_{\mathbb{1}}(x) = \delta^n \sum_{i=1}^{m} \frac{1-\kappa_i}{\kappa_i} \operatorname{grad}_z N(x,z_i) \cdot M_i \operatorname{grad} u_{\mathbb{1}}(z_i) + o(\delta^n), \quad x \in \partial\Omega \tag{95}$$

where n is again the space dimension, N the Neumann function (7), and M_i the associated polarization tensor, cf. [28] or section "Asymptotic Analysis of the Voltage Perturbations" in chapter ▸ Expansion Methods. The leading order approximation of the difference between the associated Neumann-Dirichlet operators, $\Lambda_\delta - \Lambda_{\mathbb{1}}$, can be factorized in a similar way as in Theorem 21 and has an nm-dimensional range that is spanned by dipole potentials sitting in the locations z_i of the obstacles D_i, $i = 1, \ldots, m$; recall that n is the space dimension.

For the full Maxwell's equations considered in section "Inverse Electromagnetic Scattering by an Inhomogeneous Medium," the range space of the corresponding far field operator F of (77) consists of the magnetic far fields corresponding to electric dipoles at the infinitesimal scatterers; if the scatterers also differ in their magnetic permeability, then the range space also contains the far fields of the magnetic dipoles in z_i, $i = 1, \ldots, m$.

The method described above for reconstructing the locations of small scatterers is often called *MUSIC* in the inverse problems community. Originally, the MUSIC algorithm is a signal processing tool for frequency estimation from the noisy spectrum of some signal (MUSIC stands for MUltiple SIgnal Classification.), cf., e.g., [90]. In a seminal report (Devaney, Super-resolution processing of multi-static data using time reversal and MUSIC. Unpublished manuscript, 2000), this algorithm was suggested to detect "point scatterers" on the basis of the Born approximation, which led to an algorithm that is not exactly the same, but related to the one we have sketched above. The relation between this algorithm and the factorization method has subsequently been recognized in [32, 69]. However, although the form of the factorization (94) is similar to the ones for the factorization method derived in Sects. 2 and 3, it is slightly different in its precise interpretation; this has been

exemplified in [2] by taking the limit of each of the factors from Theorem 4 as $\delta \to 0$.

The derivation of asymptotic formulas as in Theorem 20 goes back to the landmark paper [41]. In [24], formula (95) from [28] was used to provide the rigorous foundation of the MUSIC-type algorithm from above. Important extensions and generalizations to other problem settings include [1, 4, 7, 46, 91]; for a more detailed survey and further references, we refer to chapter ▸ Expansion Methods and the monographs [5, 6].

Numerical illustrations of this approach can be found in various papers (see, e.g., [3, 24, 46]).

The Singular Sources Method

As in section "Inverse Acoustic Scattering by a Sound-Soft Obstacle," we reconsider the simple inverse scattering problem for the Helmholtz equation in $\mathbb{R}^3$ to determine the shape of an acoustically soft obstacle D from the knowledge of the far field pattern $u^\infty(\hat{x};\hat{\theta})$ for all $\hat{x}, \hat{\theta} \in S^2$. We refer again to (42)–(47) for the mathematical model and the definition of the far field operator F from $L^2(S^2)$ into itself. Note that again $u^s = u^s(x;\hat{\theta})$ and $u^\infty = u^\infty(\hat{x};\hat{\theta})$ denote the scattered field and far field pattern, respectively, corresponding to the incident plane wave of direction $\hat{\theta} \in S^2$.

The basic tool in the *singular sources method* is to consider also the scattered field $v^s = v^s(x;z)$ which corresponds to the incident field $v^i(x) = \Phi(x,z)$ of (51) of a point source, where $z \notin \overline{D}$ is a given point. The scattered field $v^s(z;z)$ evaluated at the source point blows up when z tends to a boundary point. One can prove (see [73, 86]) that there exists a constant $c > 0$ (depending on D and k only) such that

$$|v^s(z;z)| \geq \frac{c}{d(z,\partial D)} \quad \text{for all } z \notin \overline{D}. \tag{96}$$

Here, $d(z,\partial D) = \inf\{|z-y| : y \in \partial D\}$ denotes the distance of z to the boundary of D.

The idea of the singular sources method is to fix $z \notin \overline{D}$ and $\varepsilon > 0$ and a bounded domain $G_z \subset \mathbb{R}^3$ such that its exterior is connected and $z \notin \overline{G_z}$ and $\overline{D} \subset G_z$. Runge's approximation theorem (see, e.g., [73]) yields the existence of $g \in L^2(S^2)$ depending on z, G_z and ε such that

$$\left\| v_g - \Phi(\cdot,z) \right\|_{C(\overline{G_z})} \leq \varepsilon, \tag{97}$$

where v_g denotes the Herglotz wave function, defined by

$$v_g(x) = \int_{S^2} g(\hat{\theta}) \exp(\mathrm{i}kx\cdot\hat{\theta}) \mathrm{d}s(\hat{\theta}), \quad x \in \mathbb{R}^3.$$

In the following, only the dependence on ε is indicated by writing g_ε. The following convergence result for the Singular Sources Method is known (see [73, 86]).

Theorem 22. *Let $u^\infty = u^\infty(\hat{x};\hat{\theta})$, $\hat{x},\hat{\theta} \in S^2$ be the far field pattern of the scattering problem (43–45). Fix $z \notin \overline{D}$ and a bounded domain $G_z \subset \mathbb{R}^3$ such that its exterior is connected and $z \notin \overline{G_z}$ and $\overline{D} \subset G_z$. For any $\varepsilon > 0$, choose $g = g_\varepsilon \in L^2(S^2)$ with (97). Then*

$$\lim_{\delta\to 0}\lim_{\varepsilon\to 0}\int_{S^2}(Fg_\varepsilon)(-\hat{\theta})g_\delta(\hat{\theta})\mathrm{d}s(\hat{\theta}) = v^s(z;z),$$

i.e., by substituting the form of F,

$$\lim_{\delta\to 0}\lim_{\varepsilon\to 0}\int_{S^2}\int_{S^2}u^\infty(-\hat{\theta};\hat{\eta})g_\varepsilon(\hat{\eta})g_\delta(\hat{\theta})\mathrm{d}s(\hat{\eta})\mathrm{d}s(\hat{\theta}) = v^s(z;z).$$

Note that the limits are iterated, i.e., first the limit w.r.t. ε has to be taken and then the limit w.r.t. δ.

Combining this result with (96) yields

$$\lim_{\delta\to 0}\lim_{\varepsilon\to 0}\left|\int_{S^2}\int_{S^2}u^\infty(-\hat{\theta};\hat{\eta})g_\varepsilon(\hat{\eta})g_\delta(\hat{\theta})\mathrm{d}s(\hat{\eta})\mathrm{d}s(\hat{\theta})\right| \geq \frac{c}{d(z,\partial D)}. \tag{98}$$

This result assures that for z sufficiently close to the boundary ∂D (and regions G_z chosen appropriately) the quantity

$$\lim_{\delta\to 0}\lim_{\varepsilon\to 0}\left|\int_{S^2}\int_{S^2}u^\infty(-\hat{\theta};\hat{\eta})g_\varepsilon(\hat{\eta})g_\delta(\hat{\theta})\mathrm{d}s(\hat{\eta})\mathrm{d}s(\hat{\theta})\right|$$

becomes large.

It is convenient to use domains G_z of the special form

$$G_{z,p} = (z+\rho p) + \left\{x \in \mathbb{R}^3 : |x| < R, \frac{x}{|x|}\cdot p > -\cos\beta\right\}$$

for some (large) radius $R > 0$, opening angle $\beta \in [0, \pi/2)$, direction of opening $p \in S^2$, and $\rho > 0$. The dependence on β, ρ, and R is not indicated since they are kept fixed. This domain $G_{z,p}$ is a ball centered at $z + \rho p$ with radius R from which the cone of direction $-p$ and opening angle β have been removed. Obviously, it is chosen such that $z \notin \overline{G_{z,p}}$. These sets $G_{z,p}$ are translations and rotations of the reference set

$$\hat{G} = \left\{x \in \mathbb{R}^3 : |x| < R, \frac{x}{|x|}\cdot \hat{p} > -\cos\beta\right\}$$

for $\hat{p} = (0, 0, 1)^{\mathrm{T}}$, i.e., $G_{z,p} = z + M\hat{G}$ for some orthogonal $M \in \mathbb{R}^{3\times 3}$.

With these transformations, we can consider the singular sources method as a sampling method with sampling objects z and M.

From the arguments used in the proof of Theorem 22, it is not clear whether or not the common limit $\lim_{\varepsilon,\,\delta\to 0}$ exists. However, if k^2 is not a Dirichlet eigenvalue of $-\Delta$ in D, then the following stronger result than (49) can be obtained by using the factorization (49).

Theorem 23. *Let $z \notin \overline{D}$ and $G_z \subset \mathbb{R}^3$ be a bounded domain such that its exterior is connected and $z \notin \overline{G_z}$ and $\overline{D} \subset G_z$. For any $\varepsilon > 0$, choose $g_\varepsilon \in L^2(S^2)$ with (97) with respect to the H^1 – norm, i.e.,*

$$\left\| v_{g_\varepsilon} - \Phi(\cdot, z) \right\|_{H^1(G_z)} \leq \varepsilon.$$

Assume furthermore that k^2 is not a Dirichlet eigenvalue of $-\Delta$ in D. Then there exists a constant $c > 0$ depending only on D and k such that

$$\left| \lim_{\varepsilon\to 0} \int_{S^2} \int_{S^2} u^\infty(\hat{\theta}; \hat{\eta}) g_\varepsilon(\hat{\eta}) \overline{g_\varepsilon(\hat{\theta})} \mathrm{d}s(\hat{\eta}) \mathrm{d}s(\hat{\theta}) \right| = \lim_{\varepsilon\to 0} \left| (Fg_\varepsilon, g_\varepsilon)_{L^2(S^2)} \right| \geq \frac{c}{d(z, \partial D)}.$$

For a proof we refer to [73]. Numerical reconstructions with the Singular Sources Methods are shown in [86].

The Probe Method

The *probe method* has originally been proposed in [62] for the inverse problem of impedance tomography of section "Conducting Obstacles," and here we also restrict our attention to this setting. To be precise, let $\sigma \in L^\infty(\Omega)$ be a (complex valued) admittivity function, and define $u \in H^1_\diamond(\Omega)$ as the unique (weak) solution of the boundary value problem

$$\mathrm{div}(\sigma \,\mathrm{grad}\, u) = 0 \quad \text{in } \Omega \qquad \sigma \frac{\partial}{\partial \nu} u = f \quad \text{on } \partial\Omega \qquad \int_{\partial\Omega} u \mathrm{d}s = 0, \tag{99}$$

where $f \in L^2_\diamond(\partial\Omega)$. For the spaces $H^1_\diamond(\Omega)$ and $L^2_\diamond(\partial\Omega)$, we refer to section "Impedance Tomography in the Presence of Insulating Inclusions."

As in section "Conducting Obstacles," we assume that $\sigma \in L^\infty(\Omega)$ is a perturbation of the constant background admittivity function $\sigma_{\mathbb{1}} = 1$. More precisely, let $D \subset \Omega$ be again the finite union of domains such that $\Omega\backslash\overline{D}$ is connected and $\sigma = 1$ in $\Omega\backslash D$, and let there be a constant $c_0 > 0$ such that

$$\mathrm{Im}\, \sigma(x) \leq 0 \;\text{ and }\; \mathrm{Re}\, \sigma(x) \geq 1 + c_0 \;\text{ on }\; D. \tag{100}$$

The case $0 < c_0 \leq \operatorname{Re} \sigma(x) \leq 1 - c_0$ can be treated in a similar way (see [73]). The unique solvability of the **direct problem**, i.e., the boundary value problem (99), guarantees existence of the *Neumann-to-Dirichlet* operators $\Lambda,\ \Lambda_{\mathbb{1}} : L^2_\diamond(\partial\Omega) \to L^2_\diamond(\partial\Omega)$ corresponding to σ and $\sigma_{\mathbb{1}} = 1$, respectively.

As in section "Conducting Obstacles," the goal of the **inverse problem** is to determine the support D of $\sigma - 1$ from the knowledge of the absolute data Λ or the relative data $\Lambda - \Lambda_{\mathbb{1}}$. The difference to the setting in section "Conducting Obstacles" is that σ is now a scalar and complex-valued function.

In the probe method, the sampling objects are curves in Ω starting at the boundary $\partial\Omega$ of Ω. In the original paper [62], these curves are called needles. We keep this notation but mention that – perhaps in contrast to the colloquial meaning – these needles do not need to be straight segments but can be curved in general. By choosing a family of needles, the probe method determines the first point on the needle which intersects the boundary ∂D (see Theorem 24 below). Therefore, in contrast to the factorization method and the linear sampling method, the probe method tests on curves instead of points.

Definition 1. A needle $\mathcal{C}$ is the image of a continuously differentiable function $\eta : [0, 1] \to \overline{\Omega}$ such that $\eta(0) \in \partial\Omega$ and $\eta(t) \in \Omega$ for all $t \in (0,\ 1]$ and $\eta'(t) \neq 0$ for all $t \in [0,\ 1]$ and $\eta(t) \neq \eta(s)$ for $t \neq s$. We call η a parameterization of the needle.

The following monotonicity property is the basic ingredient for the Probe Method.

Under the above assumptions on $\sigma \in L^\infty(\Omega)$, there exists $c > 1$ such that

$$\frac{1}{c}\int_D |\operatorname{grad} u_{\mathbb{1}}|^2 \mathrm{d}x \leq \operatorname{Re}\langle f, (\Lambda_{\mathbb{1}} - \Lambda) f\rangle \leq c \int_D |\operatorname{grad} u_{\mathbb{1}}|^2 \mathrm{d}x \tag{101}$$

for every $f \in L^2_\diamond(\partial\Omega)$. Here, $u_{\mathbb{1}} \in H^1_\diamond(\Omega)$ denotes the unique solution of (99) for the constant background case $\sigma_{\mathbb{1}} = 1$.

Let $\eta : [0, 1] \to \overline{\Omega}$ be the parameterization of a given needle, $t \in (0,\ 1]$ a fixed parameter, and $\mathcal{C}_t = \{\eta(s) : 0 \leq s \leq t\}$ the part of the needle from $s = 0$ to $s = t$. Let $\Phi(x,\ y)$ denote the fundamental solution of the Laplace equation, e.g.,

$$\Phi(x, y) = \frac{1}{4\pi|x - y|}, x \neq y,$$

in $\mathbb{R}^3$. Runge's approximation theorem (see, e.g., [73]) yields the existence of a sequence $w_n \in H^1(\Omega)$ of harmonic functions in Ω such that

$$\| w_n - \Phi(\cdot, \eta(t)) \|_{H^1(U)} \to 0,\ \ n \to \infty \tag{102}$$

for every open subset U with $\overline{U} \subset \Omega \backslash \mathcal{C}_t$. We set $f_n = \partial w_n / \partial \nu$ on $\partial\Omega$ and note that f_n depends on $\mathcal{C}_t$ but not on the unknown domain D. The dependence on $\mathcal{C}_t$ is denoted by writing $f_n(\mathcal{C}_t)$. It can – at least in principle – be computed beforehand.

Theorem 24. *Let the above assumptions on σ hold and fix a needle with parameterization $\eta : [0,1] \to \overline{\Omega}$. Define the set $\mathcal{T} \subset [0,1]$ by*

$$\mathcal{T} = \left\{ t \in [0,1] : \sup_{n \in \mathbb{N}} \{ |\mathrm{Re}\langle f_n(\mathcal{C}_t), (\Lambda - \Lambda_{\mathbb{1}}) f_n(\mathcal{C}_t) \rangle| \ < \infty \} \right. . \tag{103}$$

Here, $f_n(\mathcal{C}_t) = \partial w_n / \partial \nu \in H^{-1/2}(\partial\Omega)$ is determined from (102) (so far, we have chosen the boundary current f in (99) from $L^2_\diamond(\partial\Omega)$ for convenience; however, the quadratic form in (103) extends as dual pairing $\langle H^{-1/2}(\partial\Omega), \ H^{1/2}(\partial\Omega) \rangle$ to $f \in H^{-1/2}(\partial\Omega)$ with vanishing mean). Then $\mathcal{T} \neq \emptyset$, and one can define $t^* = \sup \{ t \in [0,1] : [0,t] \in \mathcal{T} \}$, which satisfies

$$t^* = \begin{cases} \min \{ t \in [0,1] : \eta(t) \in \partial D \}, & \text{if } \mathcal{C}_1 \ \cap \ \overline{D} \neq \emptyset \\ 1, & \text{if } \mathcal{C}_1 \ \cap \ \overline{D} \neq \emptyset. \end{cases} \tag{104}$$

We recall that $\mathcal{C}_1 = \mathcal{C} = \{ \eta(t) : t \in [0,1] \}$.

For a proof we refer to [62, 73].

Note that for every needle the set $\mathcal{T}$ of the form (103) is determined by the given data: it depends on η and the approximating functions w_n. Formula (104) provides a constructive way to determine ∂D from $\Lambda - \Lambda_{\mathbb{1}}$: one has to choose a family of needles which cover the domain Ω, and for each needle one computes t^* as the largest point of $\mathcal{T}$; if $t^* < 1$, then $\eta(t^*) \in \partial D$. Obviously, this procedure is very expensive from a computational point of view. However, if one samples with "linear" needles only, i.e., rays of the form $\mathcal{C} = \{ z + tp : t \geq 0 \} \cap \Omega$ for $z \in \Omega$ and unit vectors $p \in S^2$, then the computational effort can be reduced considerably since the approximating sequence (102) has to be computed only once for a reference needle. However, by using only rays as needles, one can not expect to detect the boundary of D completely. Only the "visible points" of ∂D can be detected, i.e., those which can be connected completely in $\Omega \backslash \overline{D}$ by straight lines to $\partial\Omega$.

In an implementation of the definition of $\mathcal{T}$ of (103), one has to decide whether a supremum is finite or infinite. Numerically, this is certainly not an easy task. In [62], it has been suggested to replace $\mathcal{T}$ of (103) by

$$\mathcal{T}_M = \{ t \in [0,1] : \sup_{n \in \mathbb{N}} \{ |\mathrm{Re}\langle f_n(\mathcal{C}_t), (\Lambda - \Lambda_{\mathbb{1}}) f_n(\mathcal{C}_t) \rangle| \leq M \}$$

for some $M > 0$, for which a result analogously to the one in Theorem 24 can be established. We refer to [62] for more details.

Again, the probe method is general enough to have extensions to a number of related inverse problems in elasticity (see [63]) and scattering theory (see [61]). For numerical reconstructions, we refer to [87].

5 Conclusion

This chapter has considered a very general approach to identify obstacles within a homogeneous background. The method requires the underlying physics to be described by an elliptic differential equation and utilizes the fact that signals become smoother the longer they travel through the homogeneous material.

This technique has been exemplified for two model problems from impedance tomography and inverse scattering. Relations of this method to other techniques such as the probe method and MUSIC have also been explored.

6 Appendix

In this appendix, we collect some functional analytic results on range identities. The factorization method makes use of the fact that the unknown domain D can be characterized by the range of some compact operator $A : X \to Y$ where A is related to the known operator $M : Y \to Y$ through the factorization

$$M = AGA^*. \tag{105}$$

Throughout this whole chapter, we assume that Y is a Hilbert space and X a reflexive Banach space with dual X^*. We denote by $A^* : Y \to X^*$ the adjoint of A, where Y is identified with its dual.

For a computable characterization of D, the range of the operator A has to be expressed by the operator M which is the goal of the *range identity*.

In the simplest case where also X is a Hilbert space and G is the identity I, the range identity is easily obtained via the singular system of A and the Theorem of Picard. We recall that $\{\sigma_j, x_j, y_j : j \in J\}$ is a singular system of a linear and compact operator $T : X \to Y$ between Hilbert spaces X and Y if $\{x_j : j \in J\}$ and $\{y_j : j \in J\}$ are complete countable orthonormal systems in the subspaces $\mathcal{N}(T)^\perp \subset X$ and $\mathcal{N}(T^*)^\perp \subset Y$, respectively, and $\sigma_j \in \mathbb{R}_{>0}$ such that $Tx_j = \sigma_j y_j$ and $T^* y_j = \sigma_j x_j$ for all $j \in J$.

We note that $\left\{\sigma_j^2, x_j : j \in J\right\}$, together with a basis of the null-space $\mathcal{N}(T)$ of T and associated eigenvalue 0, is an eigensystem of the self-adjoint and nonnegative operator T^*T. Furthermore,

$$\begin{aligned} Tx &= \sum_{j \in J} \sigma_j (x, x_j)_X y_j, \quad x \in X, \\ T^* y &= \sum_{j \in J} \sigma_j (y, y_j)_Y x_j, \quad y \in Y \end{aligned}$$

Theorem 25 (Picard). *Let X, Y be Hilbert spaces and $T : X \to Y$ be a compact operator with singular system $\{\sigma_j, x_j, y_j : j \in J\}$. Then there holds: An element $y \in Y$ belongs to the range $\mathcal{R}(T)$ of T, if and only if,*

$$y \in \mathcal{N}(T^*)^\perp \quad \text{and} \quad \sum_{j\in J} \frac{|(y, y_j)_Y|^2}{\sigma_j^2} < \infty.$$

For a proof we refer to, e.g., [39]. Applying this theorem to the factorization (105) with $G = I$, and when X^* is identified with X, one obtains.

Corollary 3. *Let $A : X \to Y$ be a compact operator between Hilbert spaces X and Y with dense range and $M = AA^* : Y \to Y$. Then the ranges of A and $M^{1/2}$ coincide. Here, the self-adjoint and nonnegative operator $M^{1/2} : Y \to Y$ is given by*

$$M^{1/2} y = \sum_{j\in J} \sqrt{\lambda_j}(y, y_j)_Y y_j, \quad y \in Y,$$

where $\{y_j : j \in J\}$ are the orthonormal eigenelements of the self-adjoint, compact, and nonnegative operator M corresponding to the positive eigenvalues λ_j. It follows that

$$y \in \mathcal{R}(A) \iff \sum_{j\in J} \frac{|(y, y_j)_Y|^2}{\lambda_j} < \infty.$$

For more general factorizations of the form $M = AGA^*$, the following (preliminary) characterization is useful (see [68]; for an equivalent formulation, see Theorem 3 of [82]).

Theorem 26. *Let X be a reflexive Banach space with dual X^* and dual form $\langle \cdot, \cdot \rangle$ in $\langle X^*, X \rangle$. Furthermore, let Y be a Hilbert space and $M : Y \to Y$ and $A : X \to Y$ be linear-bounded operators such that the factorization (105) holds for some linear and bounded operator $G : X^* \to X$, which satisfies a coercivity condition of the form: There exists $c > 0$ with*

$$|\langle \varphi, G\varphi \rangle| \geq c \|\varphi\|_{X^*}^2 \quad \text{for all } \varphi \in \mathcal{R}(A^*) \subset X^*. \tag{106}$$

Then, for any $\phi \in Y \phi \neq 0$,

$$\phi \in \mathcal{R}(A) \iff \inf\{|(\psi, M\psi)_Y| : \psi \in Y, (\psi, \phi)_Y = 1\} > 0. \tag{107}$$

Proof. The form $|(\psi, M\psi)_Y|$ can be estimated by

$$|(\psi, M\psi)_Y| = |\langle A^*\psi, GA^*\psi \rangle| \geq c \|A^*\psi\|_{X^*}^2 \quad \text{for all } \psi \in Y. \tag{108}$$

Let first $\phi = A\varphi_0$ for some $\varphi_0 \in X$. For $\psi \in Y$ with $(\psi, \phi)_Y = 1$, there holds that

$$|(\psi, M\psi)_Y| \geq c\,\|A^*\psi\|_{X^*}^2 = \frac{c}{\|\varphi_0\|_X^2}\,\|A^*\psi\|_{X^*}^2\,\|\varphi_0\|_X^2$$
$$\geq \frac{c}{\|\varphi_0\|_X^2}|\langle A^*\psi, \varphi_0\rangle|^2 = \frac{c}{\|\varphi_0\|_X^2}\big|(\psi, \underbrace{A\varphi_0}_{=\phi})_Y\big|^2 = \frac{c}{\|\varphi_0\|_X^2}.$$

This provides the lower bound of the infimum.

Second, assume that $\phi \notin \mathcal{R}(A)$. Define the closed subspace $V := \{\psi \in Y : (\psi,\ \phi)_Y = 0\}$. Then $A^*\ (V)$ is dense in $\mathcal{R}(A^*) \subset X^*$. Indeed, this is equivalent to the statement that the annihilators $[A^*(V)]^\perp$ and $[\mathcal{R}(A^*)]^\perp = \mathcal{N}(A)$ coincide. Therefore, let $\varphi \in [A^*(V)]^\perp$, i.e., $\langle A^*\psi,\ \varphi\rangle = 0$ for all $\psi \in V$, i.e., $(\psi,\ A\varphi)_Y = 0$ for all $\psi \in V$, i.e., $A\varphi \in V^\perp = \text{span}\,\{\phi\}$. Since $\phi \notin \mathcal{R}(A)$, this implies $A\varphi = 0$, i.e., $\varphi \in \mathcal{N}(A)$. Therefore, $A^*(V)$ is dense in $\mathcal{R}(A^*)$.

Choose a sequence $\{\hat\psi_n\}$ in V such that $A^*\hat\psi_n \to -\frac{1}{\|\phi\|_Y^2}A^*\phi$ as n tends to infinity and set $\psi_n = \hat\psi_n + \phi/\,\|\phi\|_Y^2$. Then $(\psi_n,\ \phi)_Y = 1$ and $A^*\psi_n \to 0$. The first equation of (108) yields

$$|(\psi_n, M\psi_n)_Y| \ \leq\ \|G\|\,\|A^*\psi_n\|_{X^*}^2$$

and thus $(\psi_n, M\psi_n)_Y \to 0, n \to \infty$, which proves that $\inf\{|(\psi, M\psi)_Y| : \psi \in Y,\ (\psi,\phi)_Y = 1\} = 0$.

We note that the inf-condition only depends on M and not on the factorization. Therefore, we have as a corollary.

Corollary 4. *Let Y be a Hilbert space and X_1 and X_2 be reflexive Banach spaces with duals X_1^* and X_2^*, respectively. Furthermore, let $M : Y \to Y$ have two factorizations of the form $M = A_1\ G_1A_1^* = A_2\ G_2A_2^*$ as in (105) with compact operators $A_j\ :\ X_j \to Y$ and bounded operators $G_j\ :\ X_j^* \to X_j$, which both satisfy the coercivity condition (106). Then the ranges of A_1 and A_2 coincide.*

Corollary 4 is useful for the analysis of the factorization method as long as M is normal. However, there are many scattering problems for which the corresponding far field operator fails to be normal, e.g., in the case of absorbing media. For these problems, one can utilize the self-adjoint operator

$$M_\# = \ |\text{Re}\ M| + \text{Im}\ M, \tag{109}$$

which can be computed from M. Note that $\text{Re}\ M = \frac{1}{2}(M + M^*)$ and $\text{Im}\ M = \frac{1}{2\text{i}}(M - M^*)$ are again self-adjoint and compact, and the absolute value $|\text{Re}\ M|$ of $\text{Re}\ M$ is defined to be

$$|\text{Re}\ M|\psi = \sum_{j\in J}|\lambda_j|(\psi, \psi_j)_Y\psi_j, \quad \psi \in Y,$$

where $\{\lambda_j, \psi_j : j \in J\}$ denotes the spectral system of $\text{Re}\ M$.

Now we can apply Corollary 4 to obtain the following result (see [73] for the lengthy proof and [76] for a weaker form of assumption (d)).

Theorem 27. *Let X be a reflexive Banach space with dual X^* and dual form $\langle \cdot, \cdot \rangle$ in $\langle X^*, X \rangle$. Furthermore, let Y be a Hilbert space and $M : Y \to Y$ and $A : X \to Y$ be linear-bounded operators such that the factorization (105) holds true for some linear and bounded operator $G : X^* \to X$. Furthermore, let the following conditions be satisfied:*

(a) *The range of A is dense in Y.*
(b) *There holds* Re $G = G_0 + G_1$, *where G_0 satisfies (106) and $G_1 : X^* \to X$ is compact.*
(c) *The imaginary part* Im G *of G is non-negative, i.e.,* $\mathrm{Im}\langle \varphi, G\varphi \rangle \geq 0$ *for all $\varphi \in X^*$.*
(d) *G is injective or* Im G *is positive on the null-space of* Re G.

Then the self-adjoint operator $M_\#$ of (109) is positive, and the ranges of A and $M_\#^{1/2}$ coincide.

As an immediate corollary, we have:

Corollary 5. *Let $M : Y \to Y$ and $A : X \to Y$ and $G : X^* \to X$ be as in Theorem 27, and let G be self-adjoint, i.e., $G^* = G$, and satisfy (106). Then the ranges of A and $M^{1/2}$ coincide, and*

$$y \in \mathcal{R}(A) \quad \Longleftrightarrow \quad \sum_{j \in J} \frac{|(y, y_j)_Y|^2}{\lambda_j} < \infty,$$

where $\{\lambda_j, y_j : j \in J\}$ denotes a spectral system of the self-adjoint and compact operator $M = AGA^*$.

Cross-References

▸ Electrical Impedance Tomography
▸ Expansion Methods
▸ Inverse Scattering

References

1. Alves, C., Ammari, H.: Boundary integral formulae for the reconstruction of imperfections of small diameter in an elastic medium. SIAM J. Appl. Math. **62**, 94–106 (2002)

2. Ammari, H., Griesmaier, R., Hanke, M.: Identification of small inhomogeneities: asymptotic factorization. Math. Comput. **76**, 1425–1448 (2007)
3. Ammari, H., Iakovleva, E., Lesselier, D.: Two numerical methods for recovering small inclusions from the scattering amplitude at a fixed frequency. SIAM J. Sci. Comput. **27**, 130–158 (2005)
4. Ammari, H., Iakovleva, E., Moskow, S.: Recovery of small inhomogeneities from the scattering amplitude at a fixed frequency. SIAM J. Math. Anal. **34**, 882–900 (2003)
5. Ammari, H., Kang, H.: Reconstruction of Small Inhomogeneities from Boundary Measurements. Lecture Notes in Mathematics, vol. 1846. Springer, New York (2004)
6. Ammari, H., Kang, H.: Polarization and Moment Tensors with Applications to Inverse Problems and Effective Medium Theory. Springer, New York (2007)
7. Ammari, H., Vogelius, M.S., Volkov, D.: Asymptotic formulas for perturbations in the electromagnetic fields due to the presence of inhomogeneities of small diameter II. The full Maxwell equations. J. Math. Pures Appl. **80**, 769–814 (2001)
8. Aramini, R., Brignone, M., Piana, M.: The linear sampling method without sampling. Inverse Probl. **22**, 2237–2254 (2006)
9. Arens, T.: Linear sampling methods for 2D inverse elastic wave scattering. Inverse Probl. **17**, 1445–1464 (2001)
10. Arens, T.: Why linear sampling works. Inverse Probl. **20**, 163–173 (2004)
11. Arens, T., Grinberg, N.I.: A complete factorization method for scattering by periodic structures. Computing **75**, 111–132 (2005)
12. Arens, T., Kirsch, A.: The factorization method in inverse scattering from periodic structures. Inverse Probl. **19**, 1195–1211 (2003)
13. Arens, T., Lechleiter, A.: The linear sampling method revisited. J. Int. Equ. Appl. **21**, 179–202 (2009)
14. Astala, K., Päivärinta, L.: Calderón's inverse conductivity problem in the plane. Ann. Math. **163**, 265–299 (2006)
15. Azzouz, M., Oesterlein, C., Hanke, M., Schilcher, K.: The factorization method for electrical impedance tomography data from a new planar device. Int. J. Biomed. Imaging **Article ID 83016**, 7p. (2007). doi:10.1155/2007/83016
16. Beretta, E., Vessella, S.: Stable determination of boundaries from Cauchy data. SIAM J. Math. Anal. **30**, 220–232 (1998)
17. Van Berkel, C., Lionheart, W.R.B.: Reconstruction of a grounded object in an electrostatic halfspace with an indicator function. Inverse Probl. Sci. Eng. **21**, 585–600 (2007)
18. Borcea, L.: Electrical impedance tomography. Inverse Probl. **18**, R99–R136 (2002)
19. Bourgeois, L., Lunéville, E.: The linear sampling method in a waveguide: a modal formulation. Inverse Probl. **24**, 015018 (2008)
20. Brignone, M., Bozza, G., Aramini, R., Pastorino, M., Piana, M.: A fully no-sampling formulation of the linear sampling method for three-dimensional inverse electromagnetic scattering problems. Inverse Probl. **25**, 015014 (2009)
21. Brühl, M.: Gebietserkennung in der elektrischen Impedanztomographie. PhD thesis, Universität Karlsruhe, Karlsruhe (1999)
22. Brühl, M.: Explicit characterization of inclusions in electrical impedance tomography. SIAM J. Math. Anal. **32**, 1327–1341 (2001)
23. Brühl, M., Hanke, M., Pidcock, M.: Crack detection using electrostatic measurements. Math. Model. Numer. Anal. **35**, 595–605 (2001)
24. Brühl, M., Hanke, M., Vogelius, M.: A direct impedance tomography algorithm for locating small inhomogeneities. Numer. Math. **93**, 635–654 (2003)
25. Burger, M., Osher, S.: A survey on level set methods for inverse problems and optimal design. Eur. J. Appl. Math. **16**, 263–301 (2005)
26. Cakoni, F., Colton, D.: The linear sampling method for cracks. Inverse Probl. **19**, 279–295 (2003)
27. Cakoni, F., Colton, D., Haddar, H.: The linear sampling method for anisotropic media. J. Comput. Appl. Math. **146**, 285–299 (2002)

28. Cedio-Fengya, D., Moskow, S., Vogelius, M.S.: Identification of conductivity imperfections of small diameter by boundary measurements. Continuous dependence and computational reconstruction. Inverse Probl. **14**, 553–595 (1998)
29. Charalambopoulos, A., Gintides, D., Kiriaki, K.: The linear sampling method for the transmission problem in three-dimensional linear elasticity. Inverse Probl. **18**, 547–558 (2002)
30. Charalambopoulos, A., Gintides, D., Kiriaki, K., Kirsch, A.: The factorization method for an acoustic wave guide. In: 7th International Workshop on Mathematical Methods in Scattering Theory and Biomedical Engineering, Nymphaio, Greece. World Scientific, Singapore, pp 120–127 (2006)
31. Charalambopoulus, A., Kirsch, A., Anagnostopoulus, K.A., Gintides, D., Kiriaki, K.: The factorization method in inverse elastic scattering from penetrable bodies. Inverse Probl. **23**, 27–51 (2007)
32. Cheney, M.: The linear sampling method and the MUSIC algorithm. Inverse Probl. **17**, 591–596 (2001)
33. Collino, F., Fares, M., Haddar, H.: Numerical and analytical study of the linear sampling method in electromagnetic inverse scattering problems. Inverse Probl. **19**, 1279–1298 (2003)
34. Colton, D., Haddar, H., Monk, P.: The linear sampling method for solving the electromagnetic inverse scattering problem. SIAM J. Sci. Comput. **24**, 719–731 (2002)
35. Colton, D., Kirsch, A.: A simple method for solving inverse scattering problems in the resonance region. Inverse Probl. **12**, 383–393 (1996)
36. Colton, D., Kress, R.: Inverse Acoustic and Electromagnetic Scattering Theory, 2nd edn. Springer, Berlin (1998)
37. Colton, D., Kress, R.: Using fundamental solutions in inverse scattering. Inverse Probl. **22**, R49–R66 (2006)
38. Colton, D., Päivärinta, L.: The uniqueness of a solution to an inverse scattering problem for electromagnetic waves. Arch. Ration. Mech. Anal. **119**, 59–70 (1992)
39. Engl, H., Hanke, M., Neubauer, A.: Regularization of Inverse Problems. Kluwer, Dordrecht (1996)
40. Fata, S.N., Guzina, B.B.: A linear sampling method for near field inverse problems in elastodynamics. Inverse Probl. **20**, 713–736 (2004)
41. Friedman, A., Vogelius, M.S.: Identification of small inhomogeneities of extreme conductivity by boundary measurements: a theorem on continuous dependence. Arch. Ration. Mech. Anal. **105**, 299–326 (1989)
42. Gebauer, B., Hyvönen, N.: Factorization method and irregular inclusions in electrical impedance tomography. Inverse Probl. **23**, 2159–2170 (2007)
43. Gebauer, B., Hyvönen, N.: Factorization method and inclusions of mixed type in an inverse elliptic boundary value problem. Inverse Probl. Imaging **2**, 355–372 (2008)
44. Gebauer, S.: The factorization method for real elliptic problems. Z. Anal. Anwend **25**, 81–102 (2006)
45. Girault, V., Raviart, P.-A.: Finite Element Methods for Navier-Stokes Equations. Springer, Berlin (1986)
46. Griesmaier, R.: An asymptotic factorization method for detecting small objects using electromagnetic scattering. SIAM J. Appl. Math. **68**, 1378–1403 (2008)
47. Griesmaier, R.: Reconstruction of thin tubular inclusions in three-dimensional domains using electrical impedance tomography. SIAM J. Imaging Sci. **3**, 340–362 (2010)
48. Grinberg, N.: Obstacle localization in an homogeneous half-space. Inverse Probl. **17**, 1113–1125 (2001)
49. Grinberg, N.: Obstacle visualization via the factorization method for the mixed boundary value problem. Inverse Probl. **18**, 1687–1704 (2002)
50. Guzina, B.B., Bonnet, M.: Topological derivative for the inverse scattering of elastic waves. Q. J. Mech. Appl. Math. **57**, 161–179 (2004)
51. Haddar, H., Monk, P.: The linear sampling method for solving the electromagnetic inverse medium problem. Inverse Probl. **18**, 891–906 (2002)
52. Hähner, P.: An inverse problem in electrostatics. Inverse Probl. **15**, 961–975 (1999)

53. Hanke, M.: Why linear sampling really seems to work. Inverse Probl. Imaging **2**, 373–395 (2008)
54. Hanke, M., Brühl, M.: Recent progress in electrical impedance tomography. Inverse Probl. **19**, S65–S90 (2003)
55. Hanke, M., Schappel, B.: The factorization method for electrical impedance tomography in the half space. SIAM J. Appl. Math. **68**, 907–924 (2008)
56. Harrach, B., Seo, J.K.: Detecting inclusions in electrical impedance tomography without reference measurements. SIAM J. Appl. Math. **69**, 1662–1681 (2009)
57. Hettlich, F.: Fréchet derivatives in inverse obstacle scattering. Inverse Probl. **11**, 371–382 (1995)
58. Hettlich, F., Rundell, W.: A second degree method for nonlinear inverse problems. SIAM J. Numer. Anal. **37**, 587–620 (2000)
59. Hyvönen, N.: Characterizing inclusions in optical tomography. Inverse Probl. **20**, 737–751 (2004)
60. Hyvönen, N.: Approximating idealized boundary data of electric impedance tomography by electrode measurements. Math. Models Methods Appl. Sci. **19**, 1185–1202 (2009)
61. Ikehata, M.: Reconstruction of an obstacle from the scattering amplitude at a fixed frequency. Inverse Probl. **14**, 949–954 (1998)
62. Ikehata, M.: Reconstruction of the shape of the inclusion by boundary measurements. Commun. Part. Diff. Equ. **23**, 1459–1474 (1998)
63. Ikehata, M.: Size estimation of inclusions. J. Inverse Ill-Posed Probl. **6**, 127–140 (1998)
64. Kaltenbacher, B., Neubauer, A., Scherzer, O.: Iterative Regularization Methods for Nonlinear Ill-Posed Problems. de Gruyter, Berlin (2008)
65. Kirsch, A.: The domain derivative and two applications in inverse scattering theory. Inverse Probl. **9**, 81–96 (1993)
66. Kirsch, A.: Characterization of the shape of a scattering obstacle using the spectral data of the far field operator. Inverse Probl. **14**, 1489–1512 (1998)
67. Kirsch, A.: Factorization of the far field operator for the inhomogeneous medium case and an application in inverse scattering theory. Inverse Probl. **15**, 413–429 (1999)
68. Kirsch, A.: New characterizations of solutions in inverse scattering theory. Appl. Anal. **76**, 319–350 (2000)
69. Kirsch, A.: The MUSIC-algorithm and the factorization method in inverse scattering theory for inhomogeneous media. Inverse Probl. **18**, 1025–1040 (2002)
70. Kirsch, A.: The factorization method for a class of inverse elliptic problems. Math. Nachr. **278**, 258–277 (2004)
71. Kirsch, A.: An integral equation for Maxwell's equations in a layered medium with an application to the factorization method. J. Int. Equ. Appl. **19**, 333–358 (2007)
72. Kirsch, A.: An integral equation for the scattering problem for an anisotropic medium and the factorization method. In: 8th International Workshop on Mathematical Methods in Scattering Theory and Biomedical Engineering, Lefkada, Greece, pp. 57–70. World Scientific, Singapore, (2007)
73. Kirsch, A., Grinberg, N.: The Factorization Method for Inverse Problems. Oxford Lecture Series in Mathematics and Its Applications, vol. 36. Oxford University Press, Oxford (2008)
74. Kirsch, A., Ritter, S.: A linear sampling method for inverse scattering from an open arc. Inverse Probl. **16**, 89–105 (2000)
75. Kress, R., Kühn, L.: Linear sampling methods for inverse boundary value problems in potential theory. Appl. Numer. Math. **43**, 161–173 (2002)
76. Lechleiter, A.: The factorization method is independent of transmission eigenvalues. Inverse Probl. Imaging **3**, 123–138 (2009)
77. Lechleiter, A., Hyvönen, N., Hakula, H.: The factorization method applied to the complete electrode model of impedance tomography. SIAM J. Appl. Math. **68**, 1097–1121 (2008)
78. Lukaschewitsch, M., Maass, P., Pidcock, M.: Tikhonov regularization for electrical impedance tomography on unbounded domains. Inverse Probl. **19**, 585–610 (2003)

79. Luke, R., Potthast, R.: The no response test – a sampling method for inverse scattering problems. SIAM J. Appl. Math. **63**, 1292–1312 (2003)
80. McLean, W.: Strongly Elliptic Systems and Boundary Integral Operators. Cambridge University Press, Cambridge (2000)
81. Monk, P.: Finite Element Methods for Maxwell's Equations. Oxford Science, Oxford (2003)
82. Nachman, A.I., Päivärinta, L., Teirilä, A.: On imaging obstacles inside inhomogeneous media. J. Funct. Anal. **252**, 490–516 (2007)
83. Pike, R., Sabatier, P.: Scattering: Scattering and Inverse Scattering in Pure and Applied Science. Academic, New York/London (2002)
84. Pironneau, O.: Optimal Shape Design for Elliptic Systems. Springer, New York (1984)
85. Potthast, R.: A fast new method to solve inverse scattering problems. Inverse Probl. **12**, 731–742 (1996)
86. Potthast, R.: Point Sources and Multipoles in Inverse Scattering Theory. Chapman & Hall/CRC, Boca Raton (2001)
87. Potthast, R.: A survey on sampling and probe methods for inverse problems. Inverse Probl. **22**, R1–R47 (2006)
88. Ringrose, J.R.: Compact Non-self-Adjoint Operators. Van Nostrand Reinhold, London (1971)
89. Sokolowski, J., Zolesio, J.P.: Introduction to Shape Optimization. Springer, New York (1992)
90. Therrien, C.W.: Discrete Random Signals and Statistical Signal Processing. Prentice-Hall, Englewood Cliffs (1992)
91. Vogelius, M.S., Volkov, D.: Asymptotic formulas for perturbations in the electromagnetic fields due to the presence of inhomogeneities of small diameter. M2AN **79**, 723–748 (2000)
92. Zou, Y., Guo, Z.: A review of electrical impedance techniques for breast cancer detection. Med. Eng. Phys. **25**, 79–90 (2003)

Inverse Scattering

David Colton and Rainer Kress

Contents

1 Introduction 650
2 Direct Scattering Problems 654
The Helmholtz Equation 654
Obstacle Scattering 657
Scattering by an Inhomogeneous Medium 660
The Maxwell Equations 661
Historical Remarks 666
3 Uniqueness in Inverse Scattering 667
Scattering by an Obstacle 667
Scattering by an Inhomogeneous Medium 669
Historical Remarks 672
4 Iterative and Decomposition Methods in Inverse Scattering 672
Newton Iterations in Inverse Obstacle Scattering 672
Decomposition Methods 675
Iterative Methods Based on Huygens' Principle 677
Newton Iterations for the Inverse Medium Problem 682
Least-Squares Methods for the Inverse Medium Problem 683
Born Approximation 685
Historical Remarks 686
5 Qualitative Methods in Inverse Scattering 686
The Far-Field Operator and Its Properties 686
The Linear Sampling Method 688
The Factorization Method 692
Lower Bounds for the Surface Impedance 693
Transmission Eigenvalues 695

D. Colton (✉)
Department of Mathematical Sciences, University of Delaware, Newark, DE, USA
e-mail: colton@math.udel.edu

R. Kress
Institut für Numerische und Angewandte Mathematik, Universität Göttingen, Göttingen, Germany
e-mail: kress@math.uni-goettingen.de

© Springer Science+Business Media New York 2015

O. Scherzer (ed.), *Handbook of Mathematical Methods in Imaging*,
DOI 10.1007/978-1-4939-0790-8_48

Historical Remarks..... 696
Cross-References..... 697
References..... 697

Abstract

We give a survey of the mathematical basis of inverse scattering theory, concentrating on the case of time-harmonic acoustic waves. After an introduction and historical remarks, we give an outline of the direct scattering problem. This is then followed by sections on uniqueness results in inverse scattering theory and iterative and decomposition methods to reconstruct the shape and material properties of the scattering object. We conclude by discussing qualitative methods in inverse scattering theory, in particular the linear sampling method and its use in obtaining lower bounds on the constitutive parameters of the scattering object.

1 Introduction

Scattering theory is concerned with the effects that obstacles and inhomogeneities have on the propagation of waves and in particular time-harmonic waves. In the context of this book, scattering theory provides the mathematical tools for imaging via acoustic and electromagnetic waves with applications to such fields as radar, sonar, geophysics, medical imaging, and nondestructive testing.

For reasons of brevity, in this survey, we focus our attention on the case of acoustic waves and only give passing references to the case of electromagnetic waves. We will furthermore give few proofs, referring the reader interested in further details to [22]. Since the literature in the area is enormous, we have only referenced a limited number of papers and hope that the reader can use these as starting point for further investigations.

Mathematical acoustics begins with the modeling of acoustic waves, i.e., sound waves. The two main media for the propagation and scattering of sound waves are air and water (underwater acoustics). A third important medium with properties close to those of water is the human body, i.e., biological tissue (ultrasound). Since sound waves are considered as small perturbations in a gas or a fluid, the equation of acoustics, i.e., the wave equation

$$\frac{1}{c^2}\,\frac{\partial^2 p}{\partial t^2} = \Delta p \tag{1}$$

for the pressure $p = p(x,t)$, is obtained by linearization of the equations for the motion of fluids. Here, $c = c(x)$ denotes the local speed of sound, and the fluid velocity is proportional to grad p. For time-harmonic acoustic waves of the form

$$p(x,t) = \mathrm{Re}\left\{u(x)\, e^{-i\omega t}\right\} \tag{2}$$

with frequency $\omega > 0$, it follows that the complex-valued space-dependent part u satisfies the reduced wave equation

$$\Delta u + \frac{\omega^2}{c^2} u = 0. \tag{3}$$

Here we emphasize that the physical quantity describing the sound wave is the real-valued sound pressure $p(x,t)$ and not the complex-valued amplitude $u(x)$ in the representation $u(x)\, e^{-i\omega t}$. For a homogeneous medium, the speed of sound c is constant and (3) becomes the *Helmholtz equation*

$$\triangle u + k^2 u = 0, \tag{4}$$

where the wave number k is given by the positive constant $k = \omega/c$.

A solution to the Helmholtz equation whose domain of definition contains the exterior of some sphere is called radiating if it satisfies the *Sommerfeld radiation condition*

$$\lim_{r\to\infty} r \left(\frac{\partial u^s}{\partial r} - iku^s \right) = 0, \tag{5}$$

where $r = |x|$ and the limit holds uniformly in all directions $x/|x|$. Here, and in the sequel, $|x| := \sqrt{x_1^2 + x_2^2 + x_3^2}$ denotes the Euclidean norm of $x = (x_1, x_2, x_3) \in \mathbb{R}^3$. For more details on the physical background of linear acoustic waves, the reader is referred to [66].

We will confine our presentation of scattering theory for time-harmonic acoustic waves to two basic problems, namely, scattering by a bounded impenetrable obstacle and scattering by a penetrable inhomogeneous medium of compact support. For a vector $d \in \mathbb{R}^3$ with $|d| = 1$, the function $e^{ik\,x\cdot d}$ satisfies the Helmholtz equation for all $x \in \mathbb{R}^3$. It is called a *plane wave*, since $e^{i(k\,x\cdot d-\omega t)}$ is constant on the planes $k\,x\cdot d - \omega t = \text{const}$. Note that these wave fronts travel with velocity c in the direction d. Assume that an incident field is given by the plane wave $u^i(x) = e^{ik\,x\cdot d}$. Then the simplest obstacle scattering problem is to find the scattered field u^s as a radiating solution to the Helmholtz equation in the exterior of a bounded scatterer D such that the total field $u = u^i + u^s$ satisfies the Dirichlet boundary condition

$$u = 0 \quad \text{on } \partial D \tag{6}$$

corresponding to a sound-soft obstacle with the total pressure, i.e., the excess pressure over the static pressure p_0, vanishing on the boundary. Concerning the geometry of scattering obstacles, for simplicity, we always will assume that D is a bounded domain with a connected boundary ∂D of class C^2. In particular, this implies that the complement $\mathbb{R}^3\backslash\overline{D}$ is connected. However, our results remain valid for a finite number of scattering obstacles.

Boundary conditions other than the Dirichlet condition also need to be considered such as the Neumann or sound-hard boundary condition

$$\frac{\partial u}{\partial \nu} = 0 \quad \text{on } \partial D \tag{7}$$

or, more generally, the impedance boundary condition

$$\frac{\partial u}{\partial \nu} + i\lambda u = 0 \quad \text{on } \partial D, \tag{8}$$

where ν is the outward unit normal to ∂D and λ is a positive constant called the surface impedance. More generally the impedance λ can also vary on ∂D. Since grad u is proportional to the fluid velocity, the impedance boundary condition describes obstacles for which the normal velocity of the fluid on the boundary is proportional to the excess pressure on the boundary. The Neumann condition corresponds to a vanishing normal velocity on the boundary. In order to avoid repetitions by considering all possible types of boundary conditions, we will in general confine ourselves to presenting the basic ideas in acoustic obstacle scattering for the case of a sound-soft obstacle.

The simplest scattering problem for an inhomogeneous medium assumes that the speed of sound is constant outside a bounded domain D. Then the total field $u = u^i + u^s$ satisfies

$$\triangle u + k^2 n u = 0 \quad \text{in } \mathbb{R}^3, \tag{9}$$

and the scattered wave u^s fulfills the Sommerfeld radiation condition (5), where the wave number is given by $k = \omega/c_0$ and $n = c_0^2/c^2$ is the *refractive index* given by the ratio of the square of the sound speeds $c = c_0$ in the homogeneous host medium and $c = c(x)$ in the inhomogeneous medium. The refractive index is positive and satisfies $n(x) = 1$ for $x \notin D$, and we assume n to be continuously differentiable in $\mathbb{R}^3$ (our results are also in general valid for n being merely piecewise continuous in $\mathbb{R}^3$). An absorbing medium is modeled by adding an absorption term which leads to a refractive index with a positive imaginary part of the form

$$n = \frac{c_0^2}{c^2} + i\,\frac{\gamma}{k}$$

in terms of a possibly space-dependent absorption coefficient γ.

Summarizing, given the incident wave and the physical properties of the scatterer, the *direct scattering problem* is to find the scattered wave and in particular its behavior at large distances from the scattering object, i.e., its far-field behavior. The *inverse scattering problem* takes this answer to the direct scattering problem as its starting point and asks what the nature of the scatterer that gave rise to such far-field behavior is.

To be more specific, it can be shown that radiating solutions u^s to the Helmholtz equation have the asymptotic behavior

$$u^s(x) = \frac{e^{ik|x|}}{|x|}\left\{u_\infty(\hat{x}) + O\left(\frac{1}{|x|}\right)\right\}, \quad |x| \to \infty, \tag{10}$$

uniformly for all directions $\hat{x} = x/|x|$, where the function u_∞ defined on the unit sphere S^2 is known as the *far-field pattern* of the scattered wave. For plane wave incidence, we indicate the dependence of the far-field pattern on the incident direction d and the observation direction $\hat{x}$ by writing $u_\infty = u_\infty(\hat{x}, d)$. The inverse scattering problem can now be formulated as the problem of determining either the sound-soft obstacle D or the index of refraction n (and hence also D) from a knowledge of the far-field pattern $u_\infty(\hat{x}, d)$ for $\hat{x}$ and d on the unit sphere S^2 (or a subset of S^2).

One of the earliest mathematical results in inverse scattering theory was Schiffer's proof in 1967 that the far-field pattern $u_\infty(\hat{x}, d)$ for $\hat{x}, d \in S^2$ uniquely determines the shape of a sound-soft obstacle D. Unfortunately, Schiffer's proof does not immediately generalize to other boundary conditions. This problem was remedied by Kirsch and Kress in 1993 who, using an idea originally proposed by Isakov, showed that $u_\infty(\hat{x}, d)$ for $\hat{x}, d \in S^2$ uniquely determines the shape of D as long as the solution of the direct scattering problem depends continuously on the boundary data [54]. In particular, it is not necessary to know the boundary condition a priori in order to guarantee uniqueness! The uniqueness problem for inverse scattering by an inhomogeneous medium was solved by Nachman [68], Novikov [70], and Ramm [79] in 1988 who based their analysis on the fundamental work of Sylvester and Uhlmann [88]. Their uniqueness proof was subsequently considerably simplified by Hähner [35].

The first attempt to reconstruct the shape of a sound-soft scattering obstacle from a knowledge of the far-field pattern in a manner acknowledging the nonlinear and ill-posed nature of the problem was made by Roger in 1981 using Newton's iteration method [81]. A characterization and rigorous proof of the existence of the Fréchet derivative of the solution operator in Newton's method were then established by Kirsch [47] and Potthast [75] in 1993 and 1994, respectively. An alternative approach to solving the inverse scattering problem was proposed by Colton and Monk in 1986 and by Kirsch and Kress in 1987 who broke up the inverse scattering problem into a linear, ill-posed problem and a nonlinear, well-posed problem [24, 53]. The optimization method of Kirsch and Kress has the attractive feature of needing only a single incident field for its implementation. On the other hand, to use such methods, it is necessary to know the number of components of the scatterer as well as the boundary condition satisfied by the field on the surface of the scatterer. These problems were overcome by Colton and Kirsch in 1996 through the derivation of a *linear* integral equation with the far-field data as its kernel (i.e., multistatic data is needed for its implementation) [19]. This method, subsequently called the *linear sampling method*, was further developed by Colton et al. [29]

and numerous other researchers. A significant development in this approach to the inverse scattering problem was the introduction of the *factorization method* by Kirsch in 1998 [48]. For further historical information on these "sampling" methods in inverse scattering theory, we refer the reader to the chapter in this handbook on sampling methods as well as the monographs [7, 52].

Optimization methods and sampling methods for the inverse scattering problem for inhomogeneous media have been extensively investigated by numerous authors. In general, the optimization methods are based on rewriting the scattering problem corresponding to (9) as the *Lippmann–Schwinger integral equation*

$$u(x) = e^{ikx\cdot d} - \frac{k^2}{4\pi}\int_{\mathbb{R}^3} \frac{e^{ik|x-y|}}{|x-y|}\, m(y)u(y)\, dy, \quad x \in \mathbb{R}^3, \tag{11}$$

where $m := 1 - n$ and the object is to determine m from a knowledge of

$$u_\infty(\hat{x}, d) = -\frac{k^2}{4\pi}\int_{\mathbb{R}^3} e^{-ik\hat{x}\cdot y} m(y)u(y)\, dy, \quad \hat{x}, d \in S^2. \tag{12}$$

On the other hand, sampling methods have also been used to study the inverse scattering problem associated with (9) where now the object is to only determine the support of m. For details and further references, see [7, 22, 52].

Finally, as pointed out in [21], an alternative direction in inverse scattering theory than that discussed above is to only try to obtain lower and upper bounds on a few relevant features of the scattering object rather than attempting a complete reconstruction. This relatively new direction in inverse scattering theory will be discussed in Sect. 5.

2 Direct Scattering Problems

The Helmholtz Equation

Most of the basic properties of solutions to the Helmholtz equation (3) can be deduced from the *fundamental solution*

$$\Phi(x, y) := \frac{1}{4\pi}\,\frac{e^{ik|x-y|}}{|x-y|}, \quad x \neq y. \tag{13}$$

For fixed $y \in \mathbb{R}^3$, it satisfies the Helmholtz equation in $\mathbb{R}^3\backslash\{y\}$. In addition, it satisfies the radiation condition (5) uniformly with respect to y on compact subsets of $\mathbb{R}^3$. Physically speaking, the fundamental solution represents an acoustic point source located at the point y. In addition to plane waves, point sources will also occur as incident fields in scattering problems.

Green's integral theorems provide basic tools for investigating the Helmholtz equation. As an immediate consequence, they imply the *Helmholtz representation*

$$u(x) = \int_{\partial D} \left\{ \frac{\partial u}{\partial \nu}(y)\, \Phi(x,y) - u(y)\, \frac{\partial \Phi(x,y)}{\partial \nu(y)} \right\} ds(y), \quad x \in D, \tag{14}$$

for solutions $u \in C^2(D) \cap C^1(\overline{D})$ to the Helmholtz equation. The representation (2) implies that solutions to the Helmholtz equation inherit analyticity from the fundamental solution. Any solution u to the Helmholtz equation in D satisfying

$$u = \frac{\partial u}{\partial \nu} = 0 \quad \text{on } \Gamma \tag{15}$$

for some open subset $\Gamma \subset \partial D$ must vanish identically in D. This can be seen via extending the definition of u by (2) for $x \in (\mathbb{R}^3 \backslash \overline{D}) \cup \Gamma$. Then, by Green's integral theorem, applied to u and $\Phi(x, \cdot)$, we have $u = 0$ in $\mathbb{R}^3 \backslash \overline{D}$. Clearly u solves the Helmholtz equation in $(\mathbb{R}^3 \backslash \partial D) \cup \Gamma$ and therefore by analyticity $u = 0$ in D since D and $\mathbb{R}^3 \backslash \overline{D}$ are connected through the gap Γ in ∂D.

As a consequence of the radiation condition (5), the Helmholtz representation is also valid in the exterior domain $\mathbb{R}^3 \backslash \overline{D}$, i.e., we have

$$u(x) = \int_{\partial D} \left\{ u(y)\, \frac{\partial \Phi(x,y)}{\partial \nu(y)} - \frac{\partial u}{\partial \nu}(y)\, \Phi(x,y) \right\} ds(y), \quad x \in \mathbb{R}^3 \backslash \overline{D}, \tag{16}$$

for radiating solutions $u \in C^2(\mathbb{R}^3 \backslash \overline{D}) \cap C^1(\mathbb{R}^3 \backslash D)$ to the Helmholtz equation. From Eq. (4) we observe that radiating solutions u to the Helmholtz equation satisfy *Sommerfeld's finiteness condition*

$$u(x) = O\left(\frac{1}{|x|} \right), \quad |x| \to \infty, \tag{17}$$

uniformly for all directions and that the validity of the Sommerfeld radiation condition (5) is invariant under translations of the origin.

We are now in a position to introduce the fundamental notion of the *far-field pattern* of radiating solutions to the Helmholtz equation.

Theorem 1. *Every radiating solution u to the Helmholtz equation has an asymptotic behavior of the form*

$$u(x) = \frac{e^{ik|x|}}{|x|} \left\{ u_\infty(\hat{x}) + O\left(\frac{1}{|x|} \right) \right\}, \quad |x| \to \infty, \tag{18}$$

uniformly in all directions $\hat{x} = x/|x|$, where the function u_∞ defined on the unit sphere S^2 is called the far-field pattern of u. Under the assumptions of (16), we have

$$u_\infty(\hat{x}) = \frac{1}{4\pi} \int_{\partial D} \left\{ u(y)\, \frac{\partial e^{-ik\hat{x}\cdot y}}{\partial \nu(y)} - \frac{\partial u}{\partial \nu}(y)\, e^{-ik\hat{x}\cdot y} \right\} ds(y), \quad \hat{x} \in S^2. \tag{19}$$

Proof. This follows from (16) by using the estimates

$$\frac{e^{ik|x-y|}}{|x-y|} = \frac{e^{ik|x|}}{|x|}\left\{e^{-ik\hat{x}\cdot y} + O\left(\frac{1}{|x|}\right)\right\}, \quad \frac{\partial}{\partial \nu(y)}\frac{e^{ik|x-y|}}{|x-y|}$$
$$= \frac{e^{ik|x|}}{|x|}\left\{\frac{\partial e^{-ik\hat{x}\cdot y}}{\partial \nu(y)} + O\left(\frac{1}{|x|}\right)\right\}$$

which hold uniformly for all $y \in \partial D$ and all directions $x/|x|$ as $|x| \to \infty$. ∎

From the representation (19), it follows that the far-field pattern is an analytic function on S^2. As an extension of (18), each radiating solution u to the Helmholtz equation has an *Atkinson–Wilcox* expansion of the form

$$u(x) = \frac{e^{ik|x|}}{|x|}\sum_{\ell=0}^{\infty}\frac{1}{|x|^{\ell}}F_{\ell}\left(\frac{x}{|x|}\right) \tag{20}$$

that converges absolutely and uniformly on compact subsets of $\mathbb{R}^3\backslash B$, where $B \supset \overline{D}$ is a ball centered at the origin. The coefficients in the expansion (20) are determined in terms of the far-field pattern $F_0 = u_\infty$ by the recursion

$$2ik\ell F_{\ell} = \ell(\ell-1)F_{\ell-1} + BF_{\ell-1}, \quad \ell = 1, 2, \ldots, \tag{21}$$

where B denotes the Laplace–Beltrami operator for the unit sphere. The following consequence of the expansion (20) is known as *Rellich's lemma*.

Lemma 1. *Let u be a radiating solution to the Helmholtz equation for which the far-field pattern u_∞ vanishes identically. Then u vanishes identically.*

Proof. This follows from (20) and (21) together with the analyticity of solutions to the Helmholtz equation. ∎

Corollary 1. *Let $u \in C^2(\mathbb{R}^3\backslash\overline{D}) \cap C^1(\mathbb{R}^3\backslash D)$ be a radiating solution to the Helmholtz equation in $\mathbb{R}^3\backslash\overline{D}$ for which*

$$\operatorname{Im}\int_{\partial D} u\frac{\partial \bar{u}}{\partial \nu}\,ds \geq 0. \tag{22}$$

Then $u = 0$ in $\mathbb{R}^3\backslash\overline{D}$.

Proof. Using Green's integral theorem, the radiation condition can be utilized to establish that

$$\lim_{r\to\infty}\int_{|x|=r}\left\{\left|\frac{\partial u}{\partial \nu}\right|^2+k^2|u|^2\right\}ds=-2k\,\mathrm{Im}\int_{\partial D}u\,\frac{\partial \bar{u}}{\partial \nu}\,ds.$$

Now the assumption (22) implies $\lim_{r\to\infty}\int_{|x|=r}|u|^2\,ds=0$, and the statement follows from Lemma 1. ■

Scattering from infinitely long cylindrical obstacles or inhomogeneities leads to the Helmholtz equation in $\mathbb{R}^2$. The two-dimensional case can be used as an approximation for scattering from finitely long cylinders, and more importantly, it can serve as a model case for testing numerical approximation schemes in direct and inverse scattering. Without giving details, we can summarize that our analysis remains valid in two dimensions after appropriate modifications of the fundamental solution and the radiation condition. The fundamental solution to the Helmholtz equation in two dimensions is given by

$$\Phi(x,y):=\frac{i}{4}\,H_0^{(1)}(k|x-y|),\quad x\neq y, \tag{23}$$

in terms of the Hankel function $H_0^{(1)}$ of the first kind of order zero. In $\mathbb{R}^2$ the Sommerfeld radiation condition has to be replaced by

$$\lim_{r\to\infty}\sqrt{r}\left(\frac{\partial u}{\partial r}-iku\right)=0,\quad r=|x|, \tag{24}$$

uniformly for all directions $x/|x|$. According to the form (24) of the radiation condition, the definition of the far-field pattern (18) has to be replaced by

$$u(x)=\frac{e^{ik|x|}}{\sqrt{|x|}}\left\{u_\infty(\hat{x})+O\left(\frac{1}{|x|}\right)\right\},\quad |x|\to\infty, \tag{25}$$

and the representation (19) assumes the form

$$u_\infty(\hat{x})=\frac{e^{i\frac{\pi}{4}}}{\sqrt{8\pi k}}\int_{\partial D}\left\{u(y)\,\frac{\partial e^{-ik\,\hat{x}\cdot y}}{\partial \nu(y)}-\frac{\partial u}{\partial \nu}\,(y)\,e^{-ik\,\hat{x}\cdot y}\right\}ds(y) \tag{26}$$

for $\hat{x}=x/|x|$.

Obstacle Scattering

After renaming the unknown functions, the direct scattering problem for sound-soft obstacles is a special case of the following exterior Dirichlet problem: Given a function $f\in C(\partial D)$, find a radiating solution $u\in C^2(\mathbb{R}^3\setminus\overline{D})\cap C(\mathbb{R}^3\setminus D)$ to the Helmholtz equation that satisfies the boundary condition

$$u = f \quad \text{on } \partial D. \tag{27}$$

Theorem 2. *The exterior Dirichlet problem for the Helmholtz equation has at most one solution.*

Proof. Let u satisfy the homogeneous boundary condition $u = 0$ on ∂D. If u were continuously differentiable up to the boundary, we could immediately apply Corollary 1 to obtain $u = 0$ in $\mathbb{R}^3 \setminus \overline{D}$. However, in the formulation of the exterior Dirichlet problem, u is only assumed to be in $C(\mathbb{R}^3 \setminus D)$. We refrain from discussing possibilities to overcome this regularity gap and refer to the literature [22]. ■

Theorem 3. *The exterior Dirichlet problem has a unique solution.*

Proof. The existence of a solution can be elegantly based on boundary integral equations. In the layer approach, one tries to find the solution in the form of a combined acoustic double- and single-layer potential

$$u(x) = \int_{\partial D} \left\{ \frac{\partial \Phi(x, y)}{\partial \nu(y)} - i \Phi(x, y) \right\} \varphi(y)\, ds(y) \tag{28}$$

for $x \in \mathbb{R}^3 \setminus \overline{D}$ with a density $\varphi \in C(\partial D)$. Then, after introducing the single- and double-layer integral operators $S, K : C(\partial D) \to C(\partial D)$ by

$$(S\varphi)(x) := 2 \int_{\partial D} \Phi(x, y) \varphi(y)\, ds(y), \quad x \in \partial D, \tag{29}$$

$$(K\varphi)(x) := 2 \int_{\partial D} \frac{\partial \Phi(x, y)}{\partial \nu(y)} \varphi(y)\, ds(y), \quad x \in \partial D, \tag{30}$$

and using their regularity and jump relations, it can be seen that (28) solves the exterior Dirichlet problem provided the density φ is a solution of the integral equation

$$\varphi + K\varphi - iS\varphi = 2f. \tag{31}$$

Due to their weakly singular kernels, the operators S and K turn out to be compact. Hence, the existence of a solution to (31) can be established with the aid of the Riesz–Fredholm theory for compact operators by showing that the homogeneous form of (31) only allows the trivial solution $\varphi = 0$.

Let φ be a solution of the homogeneous equation, and let the subscripts $\pm$ denote the limits obtained by approaching ∂D from $\mathbb{R}^3 \setminus \overline{D}$ and D, respectively. Then the potential u defined by (28) in all of $\mathbb{R}^3 \setminus \partial D$ satisfies the homogeneous boundary condition $u_+ = 0$ on ∂D whence $u = 0$ in $\mathbb{R}^3 \setminus \overline{D}$ followed by Theorem 2. The jump relations for single- and double-layer potentials now yield

$$-u_- = \varphi, \quad -\frac{\partial u_-}{\partial \nu} = i\varphi \quad \text{on } \partial D.$$

Hence, using Green's first integral theorem, we obtain

$$i \int_{\partial D} |\varphi|^2 ds = \int_{\partial D} \bar{u}_- \frac{\partial u_-}{\partial \nu}\, ds = \int_D \left\{ |\operatorname{grad} u|^2 - k^2 |u|^2 \right\} dx,$$

and taking the imaginary part yields $\varphi = 0$. ■

We note that, in addition to existence of a solution, the Riesz–Fredholm theory also establishes well-posedness, i.e., the continuous dependence of the solution on the data. Instead of the classical setting of continuous function spaces, the existence analysis can also be considered in the Sobolev space $H^{1/2}(\partial D)$ for the boundary integral operators leading to solutions in the energy space $H^1_{\text{loc}}(\mathbb{R}^3 \backslash \overline{D})$ (see [64, 69]).

We further note that without the single-layer potential included in (28), the corresponding double-layer integral equation suffers from nonuniqueness if k is a so-called irregular wave number or internal resonance, i.e., if there exist nontrivial solutions u to the Helmholtz equation in the interior domain D satisfying homogeneous Neumann boundary conditions $\partial u / \partial \nu = 0$ on ∂D.

For the numerical solution of the boundary integral equations in scattering theory via spectral methods in two and three dimensions, we refer to [22]. For boundary element methods, we refer to [80].

In general, for the scattering problem, the boundary values are as smooth as the boundary, since they are given by the restriction of the analytic function u^i to ∂D. Therefore, we may use the Helmholtz representation (16) and Green's second integral theorem applied to u^i and $\Phi(x, \cdot)$ to obtain the following theorem.

Theorem 4. *For scattering from a sound-soft obstacle D, we have*

$$u(x) = u^i(x) - \int_{\partial D} \frac{\partial u}{\partial \nu}(y)\, \Phi(x, y)\, ds(y), \quad x \in \mathbb{R}^3 \backslash \overline{D}, \tag{32}$$

and the far-field pattern of the scattered field u^s is given by

$$u_\infty(\hat{x}) = -\frac{1}{4\pi} \int_{\partial D} \frac{\partial u}{\partial \nu}(y)\, e^{-ik\hat{x}\cdot y}\, ds(y), \quad \hat{x} \in S^2. \tag{33}$$

The representation (32) for the scattered field through the so-called secondary sources on the boundary is known as *Huygens' principle*. Here we will use it for the motivation of the *Kirchhoff* or *physical optics approximation* as an intuitive procedure to simplify the direct scattering problem. For large wave numbers k, i.e., for small wave lengths, in a first approximation, a convex object D locally may be considered at each point x of ∂D as a plane with normal $\nu(x)$. This suggests

$$\frac{\partial u}{\partial \nu} = 2\,\frac{\partial u^i}{\partial \nu}$$

on the part $\partial D_- := \{x \in \partial D : \nu(x) \cdot d < 0\}$ that is illuminated and

$$\frac{\partial u}{\partial \nu} = 0$$

in the shadow region $\partial D_+ := \{x \in \partial D : \nu(x) \cdot d \geq 0\}$. Thus, the Kirchhoff approximation for the scattering of a plane wave with incident direction d at a convex sound-soft obstacle is given by

$$u(x) \approx e^{ik\,x\cdot d} - 2\int_{\partial D_-} \frac{\partial e^{ik\,y\cdot d}}{\partial \nu(y)}\,\Phi(x,y)\,ds(y) \tag{34}$$

for $x \in \mathbb{R}^3 \backslash \overline{D}$.

Scattering by an Inhomogeneous Medium

Recall the scattering problem for an inhomogeneous medium with refractive index n as described by (9) for the total wave $u = u^i + u^s$ with incident field u^i and the scattered wave u^s satisfying the Sommerfeld radiation condition. The function $m := 1 - n$ has support $\overline{D}$.

The counterpart of the Helmholtz representation is given by the *Lippmann–Schwinger equation*

$$u(x) = u^i(x) - k^2 \int_D \Phi(x,y) m(y) u(y)\,dy, \quad x \in \mathbb{R}^3, \tag{35}$$

which can be shown to be equivalent to the scattering problem.

In order to establish existence of a solution to (35) via the Riesz–Fredholm theory, it must be shown that the homogeneous equation has only the trivial solution or, equivalently, that the only solution to (9) satisfying the radiation condition is identically zero. For this, in addition to Rellich's lemma, the following *unique continuation principle* is required: Any solution $u \in C^2(G)$ of Eq. (9) in a domain $G \subset \mathbb{R}^3$ such that $n \in C(G)$ and u vanishes in an open subset of G vanishes identically. Hence, we have the following result on existence and uniqueness for the inhomogeneous medium scattering problem.

Theorem 5. *For a refractive index $n \in C^1(\mathbb{R}^3)$ with $\operatorname{Re} n \geq 0$ and $\operatorname{Im} n \geq 0$, the Lippmann–Schwinger equation or, equivalently, the inhomogeneous medium scattering problem has a unique solution.*

Proof. From Green's first integral theorem, it follows that

$$\int_{\partial D} u \frac{\partial \bar{u}}{\partial \nu}\, ds = \int_D \left\{ |\operatorname{grad} u|^2 - k^2 \overline{n} |u|^2 \right\} dx.$$

Taking the imaginary part and applying Corollary 1 yields $u = 0$ in $\mathbb{R}^3 \setminus D$ in view of the assumptions on n, and the proof is finished by the unique continuation principle. ■

From (35) we see that

$$u^s(x) = -k^2 \int_{\mathbb{R}^3} \Phi(x, y) m(y) u(y)\, dy, \quad x \in \mathbb{R}^3.$$

Hence, the far-field pattern u_∞ is given by

$$u_\infty(\hat{x}) = -\frac{k^2}{4\pi} \int_{\mathbb{R}^3} e^{-ik\hat{x}\cdot y} m(y) u(y)\, dy, \quad \hat{x} \in S^2. \tag{36}$$

We note that for $k^2 \|m\|_\infty$ sufficiently small, u can be obtained by the method of successive approximations. If in (36) we replace u by the first term in this iterative process, we obtain the *Born approximation*

$$u_\infty(\hat{x}) \approx -\frac{k^2}{4\pi} \int_{\mathbb{R}^3} e^{-ik\hat{x}\cdot y} m(y) u^i(y)\, dy, \quad \hat{x} \in S^2. \tag{37}$$

For numerical solutions of the inverse medium scattering problem by finite element methods coupled with boundary element methods via nonlocal boundary conditions, we refer to [65].

The Maxwell Equations

We now consider the *Maxwell equations* as the foundation of electromagnetic scattering theory. Our presentation is organized parallel to that on the Helmholtz equation, i.e., on acoustic scattering, and will be confined to homogeneous isotropic media. Consider electromagnetic wave propagation in an isotropic dielectric medium in $\mathbb{R}^3$ with constant electric permittivity ε and magnetic permeability μ. The electromagnetic wave is described by the electric field $\mathcal{E}$ and the magnetic field $\mathcal{H}$ satisfying the time-dependent Maxwell equations

$$\operatorname{curl} \mathcal{E} + \mu \frac{\partial \mathcal{H}}{\partial t} = 0, \quad \operatorname{curl} \mathcal{H} - \varepsilon \frac{\partial \mathcal{E}}{\partial t} = 0. \tag{38}$$

For time-harmonic electromagnetic waves of the form

$$\mathcal{E}(x, t) = \operatorname{Re}\left\{ \varepsilon^{-1/2} E(x)\, e^{-i\omega t} \right\}, \quad \mathcal{H}(x, t) = \operatorname{Re}\left\{ \mu^{-1/2} H(x)\, e^{-i\omega t} \right\} \tag{39}$$

with frequency $\omega > 0$, the complex-valued space-dependent parts E and H satisfy the reduced Maxwell equations

$$\operatorname{curl} E - ikH = 0, \quad \operatorname{curl} H + ikE = 0, \tag{40}$$

where the wave number k is given by the positive constant $k = \sqrt{\varepsilon\mu}\,\omega$. We will only be concerned with the reduced Maxwell equations and will henceforth refer to them as the Maxwell equations.

A solution E, H to the Maxwell equations whose domain of definition contains the exterior of some sphere is called radiating if it satisfies one of the *Silver–Müller radiation conditions*

$$\lim_{r\to\infty} (H \times x - rE) = 0 \tag{41}$$

or

$$\lim_{r\to\infty} (E \times x + rH) = 0, \tag{42}$$

where $r = |x|$ and the limits hold uniformly in all directions $x/|x|$. For more details on the physical background of electromagnetic waves, we refer to [46, 67].

For the Maxwell equations, the counterpart of the Helmholtz representation (2) is given by the *Stratton–Chu formula*

$$\begin{aligned} E(x) = &- \operatorname{curl} \int_{\partial D} \nu(y) \times E(y)\, \Phi(x, y)\, ds(y) \\ &+ \frac{1}{ik} \operatorname{curl}\operatorname{curl} \int_{\partial D} \nu(y) \times H(y)\, \Phi(x, y)\, ds(y), \quad x \in D, \end{aligned} \tag{43}$$

for solutions $E, H \in C^1(D) \cap C(\overline{D})$ to the Maxwell equations. A corresponding representation for H can be obtained from (43) with the aid of $H = \operatorname{curl} E/ik$.

The representation (43) implies that each continuously differentiable solution to the Maxwell equations automatically has analytic Cartesian components. Therefore, one can employ the vector identity $\operatorname{curl}\operatorname{curl} E = -\Delta E + \operatorname{grad}\operatorname{div} E$ to prove that for a solution E, H to the Maxwell equations, both E and H are divergence-free and satisfy the vector Helmholtz equation. Conversely, if E is a solution to the vector Helmholtz equation $\Delta E + k^2 E = 0$ satisfying $\operatorname{div} E = 0$, then E and $H := \operatorname{curl} E/ik$ satisfy the Maxwell equations.

It can be shown that solutions E, H to the Maxwell equations in D satisfying

$$\nu \times E = \nu \times H = 0 \quad \text{on } \Gamma \tag{44}$$

for some open subset $\Gamma \subset \partial D$ must vanish identically in D.

As a consequence of the Silver–Müller radiation condition, the Stratton–Chu formula is also valid in the exterior domain $\mathbb{R}^3 \setminus \overline{D}$, i.e., we have

$$E(x) = \operatorname{curl} \int_{\partial D} \nu(y) \times E(y)\, \Phi(x,y)\, ds(y) \\ -\frac{1}{ik} \operatorname{curl}\operatorname{curl} \int_{\partial D} \nu(y) \times H(y)\, \Phi(x,y)\, ds(y), \quad x \in \mathbb{R}^3 \backslash \overline{D}, \tag{45}$$

for radiating solutions $E, H \in C^1(\mathbb{R}^3 \backslash \overline{D}) \cap C(\mathbb{R}^3 \backslash D)$ to the Maxwell equations. Again, a corresponding representation for H can be obtained from (45) with the aid of $H = \operatorname{curl} E/ik$.

From (45) it can be seen that the radiation condition (41) implies (42) and vice versa. Furthermore, one can deduce that radiating solutions E, H to the Maxwell equations automatically satisfy the *Silver–Müller finiteness conditions*

$$E(x) = O\left(\frac{1}{|x|}\right), \quad H(x) = O\left(\frac{1}{|x|}\right), \quad |x| \to \infty, \tag{46}$$

uniformly for all directions and that the validity of the Silver–Müller radiation conditions (41) and (42) is invariant under translations of the origin. From the Helmholtz representation (16) for radiating solutions to the Helmholtz equation and the Stratton–Chu formulas for radiating solutions to the Maxwell equations, it can be deduced that for solutions to the Maxwell equations, the Silver–Müller radiation condition is equivalent to the Sommerfeld radiation condition for the Cartesian components of E and H.

The Stratton–Chu formula (45) can be used to introduce the notion of the *electric* and *magnetic far-field patterns*.

Theorem 6. *Every radiating solution E, H to the Maxwell equations has the asymptotic form*

$$E(x) = \frac{e^{ik|x|}}{|x|} \left\{ E_\infty(\hat{x}) + O\left(\frac{1}{|x|}\right) \right\}, \quad H(x) = \frac{e^{ik|x|}}{|x|} \left\{ H_\infty(\hat{x}) + O\left(\frac{1}{|x|}\right) \right\} \tag{47}$$

for $|x| \to \infty$ uniformly in all directions $\hat{x} = x/|x|$, where the vector fields E_∞ and H_∞ defined on the unit sphere S^2 are called the electric far-field pattern and magnetic far-field pattern, respectively. They satisfy

$$H_\infty = \nu \times E_\infty \quad \text{and} \quad \nu \cdot E_\infty = \nu \cdot H_\infty = 0 \tag{48}$$

with the unit outward normal ν on S^2. Under the assumptions of (45), we have

$$E_\infty(\hat{x}) = \frac{ik}{4\pi}\, \hat{x} \times \int_{\partial D} \{\nu(y) \times E(y) + [\nu(y) \times H(y)] \times \hat{x}\}\, e^{-ik\,\hat{x}\cdot y}\, ds(y) \tag{49}$$

for $\hat{x} \in S^2$ and a corresponding expression for H_∞.

Rellich's lemma carries over immediately from the Helmholtz to the Maxwell equations.

Lemma 2. *Let E, H be a radiating solution to the Maxwell equations for which the electric far-field pattern E_∞ vanishes identically. Then E and H vanish identically.*

The electromagnetic counterpart of Corollary 1 is given by the following result.

Corollary 2. *Let $E, H \in C^1(\mathbb{R}^3\setminus\overline{D}) \cap C(\mathbb{R}^3\setminus D)$ be a radiating solution to the Maxwell equations in $\mathbb{R}^3\setminus\overline{D}$ for which*

$$\operatorname{Re}\int_{\partial D} \nu\cdot E\times\bar{H}\,ds\le 0.$$

Then $E = H = 0$ in $\mathbb{R}^3\setminus\overline{D}$.

For two vectors $d, p \in \mathbb{R}^3$ with $|d| = 1$ and $p\cdot d = 0$, the *plane waves*

$$E^i(x) = p\,e^{ik\,x\cdot d}, \quad H^i(x) = d\times p\,e^{ik\,x\cdot d} \tag{50}$$

satisfy the Maxwell equations for all $x \in \mathbb{R}^3$. The orthogonal vectors p and $d \times p$ describe the polarization direction of the electric and the magnetic field, respectively. Given the incident field E^i, H^i and a bounded domain $D \subset \mathbb{R}^3$, the simplest obstacle scattering problem is to find the scattered field E^s, H^s as a radiating solution to the Maxwell equations in the exterior of the scatterer D such that the total field $E = E^i + E^s$, $H = H^i + H^s$ satisfies the perfect conductor boundary condition

$$\nu\times E = 0 \quad \text{on } \partial D, \tag{51}$$

where ν is the outward unit normal to ∂D. A more general boundary condition is the impedance or Leontovich boundary condition

$$\nu\times H - \lambda\,(\nu\times E)\times\nu = 0 \quad \text{on } \partial D, \tag{52}$$

where λ is a positive constant or function called the surface impedance.

Theorem 7. *The scattering problem for a perfect conductor has a unique solution.*

Proof. Uniqueness follows from Corollary 2. The existence of a solution again can be based on boundary integral equations. In the layer approach, one tries to find the solution in the form of a combined magnetic and electric dipole distribution

$$\begin{aligned} E^s(x) = \operatorname{curl}\int_{\partial D} a(y)\Phi(x,y)\,ds(y) \\ +i\,\operatorname{curl}\operatorname{curl}\int_{\partial D} \nu(y)\times\left(S_0^2 a\right)(y)\Phi(x,y)\,ds(y), \quad x\in\mathbb{R}^3\setminus\partial D. \end{aligned} \tag{53}$$

Here S_0 denotes the single-layer operator (29) in the potential theoretic limit case $k = 0$, and the density a is assumed to belong to the space $C^{0,\alpha}_{\mathrm{div}}(\partial D)$ of Hölder continuous tangential fields with Hölder continuous surface divergence. After defining the electromagnetic boundary integral operators M and N by

$$(Ma)(x) := 2\int_{\partial D} \nu(x) \times \mathrm{curl}_x \{a(y)\Phi(x,y)\}\, ds(y), \quad x \in \partial D, \tag{54}$$

and

$$(Na)(x) := 2\,\nu(x) \times \mathrm{curl}\,\mathrm{curl}\int_{\partial D} \nu(y) \times a(y)\,\Phi(x,y)\, ds(y), \quad x \in \partial D, \tag{55}$$

it can be shown that E^s given by (53) together with $H^s = \mathrm{curl}\, E^s/ik$ solves the perfect conductor scattering problem provided the density a satisfies the integral equation

$$a + Ma + iNPS_0^2 a = -2\,\nu \times E^i. \tag{56}$$

Here the operator P is defined by $Pb := (\nu \times b) \times \nu$ for (not necessarily tangential) vector fields b. Exploiting the smoothing properties of the operator S_0, it can be shown that $M + iNPS_0^2$ is a compact operator from $C^{0,\alpha}_{\mathrm{div}}(\partial D)$ into itself. The existence of a solution to (56) can now be based on the Riesz–Fredholm theory by establishing that the homogeneous form of (56) only has the trivial solution [22]. ■

Note that, analogous to the acoustic case, without the electric dipole distribution included in (53), the corresponding magnetic dipole integral equation is not uniquely solvable if k is a irregular wave number, i.e., if there exists a nontrivial solution E, H to the Maxwell equations in D satisfying the homogeneous boundary condition $\nu \times E = 0$ on ∂D.

Instead of the classical setting of Hölder continuous functions, the integral equation can also be considered in the Sobolev space $H^{1/2}_{\mathrm{div}}(\partial D)$ of tangential fields in $H^{1/2}(\partial D)$ that have a weak surface divergence in $H^{1/2}(\partial D)$ (see [69]).

In addition to electromagnetic obstacle scattering, one can also consider scattering from an inhomogeneous medium where outside a bounded domain D the electric permittivity ε and magnetic permeability μ are constant and the conductivity σ vanishes, i.e., $\varepsilon = \varepsilon_0$, $\mu = \mu_0$ and $\sigma = 0$ in $\mathbb{R}^3 \setminus \overline{D}$. For simplicity we will assume that the magnetic permeability is constant throughout $\mathbb{R}^3$. Then, again assuming the time-harmonic form (39) with ε and μ replaced by ε_0 and μ_0, respectively, the total fields $E = E^i + E^s$, $H = H^i + H^s$ satisfy

$$\mathrm{curl}\, E - ikH = 0, \quad \mathrm{curl}\, H + iknE = 0 \quad \text{in } \mathbb{R}^3 \tag{57}$$

and the scattered wave E^s, H^s satisfies the Silver–Müller radiation condition, where the wave number is given by $k = \sqrt{\varepsilon_0\mu_0}\,\omega$ and $n = (\varepsilon + i\,\sigma/\omega)/\varepsilon_0$ is the refractive index. Establishing uniqueness requires an electromagnetic analogue of the unique continuation principle, and existence can be based on an electromagnetic variant of the Lippmann–Schwinger equation [22].

The scattering of time-harmonic electromagnetic waves by an infinitely long cylinder with a simply connected cross section D leads to boundary value problems for the two-dimensional Helmholtz equation in the exterior $\mathbb{R}^2\backslash\overline{D}$ of D. If the electric field is polarized parallel to the axis of the cylinder and if the axis of the cylinder is parallel to the x_3 axis, then

$$E = (0, 0, u), \quad H = \frac{1}{ik}\left(\frac{\partial u}{\partial x_2}, -\frac{\partial u}{\partial x_1}, 0\right)$$

satisfies the Maxwell equations if and only if $u = u(x_1, x_2)$ satisfies the Helmholtz equation. The homogeneous perfect conductor boundary condition is satisfied on the boundary of the cylinder if the homogeneous Dirichlet boundary condition $u = 0$ on ∂D is satisfied. If the magnetic field is polarized parallel to the axis of the cylinder, then the roles of E and H have to be reversed, i.e.,

$$H = (0, 0, u), \quad E = \frac{i}{k}\left(\frac{\partial u}{\partial x_2}, -\frac{\partial u}{\partial x_1}, 0\right),$$

and the perfect conductor boundary condition corresponds to the Neumann boundary condition $\partial u/\partial \nu = 0$ on ∂D with the unit normal ν to the boundary ∂D of the cross section D. Hence, the analysis of two-dimensional electromagnetic scattering problems coincides with that of two-dimensional acoustic scattering problems.

Historical Remarks

Equation (4) carries the name of Helmholtz (1821–1894) for his contributions to mathematical acoustics. The radiation condition (5) was introduced by Sommerfeld in 1912 to characterize an outward energy flux. The expansion (20) was first established by Atkinson in 1949 and generalized by Wilcox in 1956. The fundamental Lemma 1 is due to Rellich (1943) and Vekua (1943). The combined single- and double-layer approach (28) for the existence analysis was introduced independently by Leis, Brakhage and Werner, and Panich in the 1960s in order to remedy the nonuniqueness deficiency of the classical double-layer approach due to Vekua, Weyl, and Müller from the 1950s. Huygens' principle is also referred to as the Huygens–Fresnel principle and named for Huygens (1629–1695) and Fresnel (1788–1827) in recognition of their contributions to wave optics. The physical optics approximation (34) is named for Kirchhoff (1824–1887) for his contributions to optics. The terms Lippmann–Schwinger equation and Born approximation are adopted from quantum physics. Equation (38) is named for Maxwell (1831–

1879) for his fundamental contributions to electromagnetic theory. The radiation conditions (41) and (42) were independently introduced in the 1940s by Silver and Müller. The integral representations (43) and (45) were first presented by Stratton and Chu in 1939. Extending the ideas of Leis, Brakhage and Werner, and Panich from acoustics to electromagnetics, the approach (53) was introduced independently by Knauff and Kress, Jones, and Mautz and Harrington in the 1970s in order to remedy the nonuniqueness deficiency of the classical approach due to Weyl and Müller from the 1950s.

3 Uniqueness in Inverse Scattering

Scattering by an Obstacle

The first step in studying any inverse scattering problem is to establish a uniqueness result, i.e., if a given set of data is known exactly, does this data uniquely determine the support and/or the material properties of the scatterer? We will begin with the case of scattering by an impenetrable obstacle and then proceed to the case of a penetrable obstacle.

From section "Obstacle Scattering," we recall that the direct obstacle scattering problem is to find $u \in C^2(\mathbb{R}^3\backslash\overline{D}) \cap C(\mathbb{R}^3\backslash D)$ such that the total field $u = u^i + u^s$ satisfies the Helmholtz equation

$$\Delta u + k^2 u = 0 \quad \text{in } \mathbb{R}^3\backslash\overline{D} \tag{58}$$

and the sound-soft boundary condition

$$u = 0 \quad \text{on } \partial D, \tag{59}$$

where $u^i(x) = e^{ikx\cdot d}$, $|d| = 1$, and u^s is a radiating solution. We also recall from Theorem 1 that u^s has the asymptotic behavior

$$u^s(x,d) = \frac{e^{ik|x|}}{|x|}\left\{u_\infty(\hat{x},d) + O\left(\frac{1}{|x|}\right)\right\}, \quad |x| \to \infty, \tag{60}$$

uniformly for all directions $\hat{x} = x/|x|$, where u_∞ is the far-field pattern of the scattered field E^s. By Green's integral theorem and the far-field representation (33), it can be shown that the far-field pattern satisfies the *reciprocity relation* [22]

$$u_\infty(\hat{x},d) = u_\infty(-d,-\hat{x}), \quad \hat{x},d \in S^2. \tag{61}$$

The *inverse scattering problem* we are concerned with is to determine D from a knowledge of $u_\infty(\hat{x},d)$ for $\hat{x}$ and d on the unit sphere S^2 and fixed wave number k. In particular, for the acoustic scattering problem (58) and (59), we want to show that D is uniquely determined by $u_\infty(\hat{x},d)$ for $\hat{x},d \in S^2$. We note that by the

reciprocity relation (61), the far-field pattern u_∞ is an analytic function of both $\hat{x}$ and d, and hence it would suffice to consider u_∞ for $\hat{x}$ and d restricted to a surface patch of the unit sphere S^2.

Theorem 8. *Assume that D_1 and D_2 are two obstacles such that the far-field patterns corresponding to the exterior Dirichlet problem (58) and (59) for D_1 and D_2 coincide for all incident directions d. Then $D_1 = D_2$.*

Proof. Let u_1^s and u_2^s be the scattered fields corresponding to D_1 and D_2, respectively. By the analyticity of the scattered field as a function of x and Rellich's Lemma 1, the scattered fields satisfy $u_1^s(\cdot, d) = u_2^s(\cdot, d)$ in the unbounded component G of the complement of $\overline{D_1} \cup \overline{D_2}$ for all $d \in S^2$. This in turn implies that the scattered fields corresponding to $\Phi(\cdot, z)$ as incident field and D_1 or D_2 as the scattering obstacle satisfy $u_1^s(x, z) = u_2^s(x, z)$ for all $x, z \in G$. Now assume that $D_1 \neq D_2$. Then, without loss of generality, there exists $x^* \in \partial G$ such that $x^* \in \partial D_1$ and $x^* \notin \overline{D_2}$. Then setting $z_n := x^* + \frac{1}{n}\nu(x^*)$, we have that $\lim_{n\to\infty} u_2^s(x^*, z_n)$ exists but $\lim_{n\to\infty} u_1^s(x^*, z_n) = \infty$ which is a contradiction and hence $D_1 = D_2$. ∎

An open problem is to determine if one incident plane wave at a fixed wave number k is sufficient to uniquely determine the scatterer D. If it is known a priori that in addition to the sound-soft boundary condition (59) D is contained in a ball of radius R such that $kR < 4.49$, then D is uniquely determined by its far-field pattern for a single incident direction d and fixed wave number k [32] (see also [22]). D is also uniquely determined if instead of assuming that D is contained in a ball of sufficiently small radius, it is assumed that D is close to a given obstacle [87]. It is also known that for a wide class of sound-soft scatterers, a finite number of incident fields are sufficient to uniquely determine D [82]. Finally, if it is assumed that D is polyhedral, then a single incident plane wave is sufficient to uniquely determine D [1, 63].

We conclude this section on uniqueness results for the inverse scattering problem for an obstacle by considering the scattering of electromagnetic waves by a perfectly conducting obstacle D. From section "The Maxwell Equations" we recall that the direct obstacle scattering problem is to find $E, H \in C^2(\mathbb{R}^3\setminus\overline{D}) \cap C(\mathbb{R}^3\setminus D)$ such that the total field $E = E^i + E^s$, $H = H^i + H^s$ satisfies the Maxwell equations

$$\operatorname{curl} E - ikH = 0, \quad \operatorname{curl} H + ikE = 0 \quad \text{in } \mathbb{R}^3\setminus\overline{D} \tag{62}$$

and the perfect conductor boundary condition

$$\nu \times E = 0 \quad \text{on } \partial D, \tag{63}$$

where E^i, H^i is the plane wave given by (50) and E^s, H^s is a radiating solution. We also recall from Theorem 1 that u^s has the asymptotic behavior

$$E^s(x,d,p) = \frac{e^{ik|x|}}{|x|}\left\{E_\infty(\hat{x},d,p) + O\left(\frac{1}{|x|}\right)\right\}, \quad |x| \to \infty, \tag{64}$$

where E_∞ is the electric far-field pattern of the scattered field E^s.

The inverse scattering problem is to determine D from a knowledge of $E_\infty(\hat{x},d,p)$ for $\hat{x}$ and d on the unit sphere S^2, three linearly independent polarizations p, and fixed wave number k. We note that E_∞ is an analytic function of $\hat{x}$ and d and is linear with respect to p. The following theorem can be proved using the same ideas as in the proof of Theorem 8.

Theorem 9. *Assume that D_1 and D_2 are two perfect conductors such that for a fixed wave number k, the electric far-field patterns for both scatterers coincide for all incident directions d and three linearly independent polarizations p. Then $D_1 = D_2$.*

In the case when D consists of finitely many polyhedra, a single incident wave is sufficient to uniquely determine D [62].

Scattering by an Inhomogeneous Medium

We now return to scattering of acoustic waves, but instead of scattering by a sound-soft obstacle, we consider scattering by an inhomogeneous medium where the governing equation (see section "Scattering by an Inhomogeneous Medium") is

$$\Delta u + k^2 n u = 0 \quad \text{in } \mathbb{R}^3 \tag{65}$$

for $u = u^i + u^s \in C^2(\mathbb{R}^2)$, where $n \in C^1(\mathbb{R}^3)$ is the refractive index satisfying $\operatorname{Re} n > 0$ and $\operatorname{Im} n \geq 0$, $u^i(x) = e^{ikx\cdot d}$ and u^s is radiating. We let $\overline{D}$ denote the support of $m := 1 - n$. By Theorem 1 the scattered wave u^s again has the asymptotic behavior (60). The inverse scattering problem we are now concerned with is to determine the index of refraction n (and hence D) from a knowledge of $u_\infty(\hat{x},d)$ for $\hat{x}$ and d on the unit sphere S^2 and fixed wave number k. In particular, we want to show that n is uniquely determined from $u_\infty(\hat{x},d)$ for $\hat{x}, d \in S^2$ and fixed wave number k.

Theorem 10. *The refractive index n in (65) is uniquely determined by $u_\infty(\hat{x},d)$ for $\hat{x}, d \in S^2$ and a fixed value of the wave number k.*

Proof. Let B be an open ball centered at the origin and containing the support of $m = 1 - n$. The first step in the proof is to construct a solution of (65) in B of the form

$$w(x) = e^{iz\cdot x}(1 + r(x)), \tag{66}$$

where $z \cdot z = 0$, $z \in \mathbb{C}^3$, and

$$\|r\|_{L^2(B)} \leq \frac{C}{|\operatorname{Re} z|}$$

for some positive constant C and $|\operatorname{Re} z|$ sufficiently large. This is done in [35] by using Fourier series. The second step is to show that, given two open balls B_1 and B_2 centered at the origin and containing the support of m such that $\overline{B_1} \subset B_2$, the set of solutions $\{u(\cdot, d) : d \in S^2\}$ satisfying (65) is complete in

$$H := \{w \in C^2(B_2) : \Delta w + k^2 n w = 0 \quad \text{in } B_2\}$$

with respect to the norm in $L^2(B_1)$ [35]. Now assume that n_1 and n_2 are refractive indices such that the corresponding far-field patterns satisfy $u_{1,\infty}(\cdot, d) = u_{2,\infty}(\cdot, d)$, $d \in S^2$, and assume that the supports of $1 - n_1$ and $1 - n_2$ are contained in $\overline{B_1}$. Then using Rellich's Lemma 1 and Green's integral theorem, it can be shown that

$$\int_{B_1} u_1(\cdot, \tilde{d}) u_2(\cdot, d)(n_1 - n_2)\, dx = 0$$

for all $d, \tilde{d} \in S^2$ and hence

$$\int_{B_1} w_1 w_2 (n_1 - n_2)\, dx = 0 \tag{67}$$

for all solutions $w_1, w_2 \in C^2(B_2)$ of $\Delta w_1 + k^2 n_1 w = 0$ and $\Delta w_2 + k^2 n_2 w_2 = 0$ in B_2. Now choose $z_1 := y + \rho a + ib$ and $z_2 := y - \rho a - ib$ such that $\{y, a, b\}$ is an orthogonal basis in $\mathbb{R}^3$ with the properties that $|a| = 1$ and $|b|^2 = |y|^2 + \rho^2$, and substitute these values of z into (66) arriving at functions w_1 and w_2. Substitute these functions into (67), and let $\rho \to \infty$ to arrive at

$$\int_{B_1} e^{2i\, y \cdot x} \left(n_1(x) - n_2(x)\right)\, dx = 0$$

for arbitrary $y \in \mathbb{R}^3$, i.e., $n_1(x) = n_2(x)$ for $x \in B_1$ by the Fourier integral theorem. ■

In the case of scattering by a sound-soft obstacle, the proof of uniqueness given in Theorem 8 remains valid in $\mathbb{R}^2$. However, this is not the case for scattering by an inhomogeneous medium. Indeed, until recently, the question of whether or not Theorem 10 remains valid in $\mathbb{R}^2$ was one of the outstanding open problems in inverse scattering theory. The problem was finally resolved in 2008 by Bukhgeim [4] who showed that in $\mathbb{R}^2$ the index of refraction n is uniquely determined by $u_\infty(\hat{x}, d)$ for $\hat{x}, d \in S^1$ and a fixed value of the wave number k.

We conclude this section with a few remarks on scattering by an anisotropic medium. Let n be as above and recall that $\overline{D}$ is the support of $m := 1 - n$. Let A be a 3×3 matrix-valued function whose entries a_{jk} are continuously differentiable functions in $\overline{D}$ such that A is symmetric and satisfies

$$\overline{\xi} \cdot (\operatorname{Im} A)\xi \leq 0, \quad \overline{\xi} \cdot (\operatorname{Re} A)\xi > \gamma |\xi|^2$$

for all $\xi \in \mathbb{C}^3$ and $x \in D$, where γ is a positive constant. We assume that $A(x) = I$ for $x \in \mathbb{R}^3 \backslash \overline{D}$. The anisotropic scattering problem is to find $u = u^i + u^s \in H^1_{\text{loc}}(\mathbb{R}^3)$ such that

$$\nabla \cdot A \nabla u + k^2 n u = 0 \quad \text{in } \mathbb{R}^3 \tag{68}$$

in the weak sense where again $u^i(x) = e^{ikx \cdot d}$ and u^s is radiating. The existence of a unique solution to this scattering problem has been established by Hähner [36].

The scattered field again has the asymptotics (60). The inverse scattering problem is now to determine D from a knowledge of the far-field pattern $u_\infty(\hat{x}, d)$ for $\hat{x}, d \in S^2$. We note that the matrix A is not uniquely determined by u_∞, and hence without further a priori assumptions, the determination of D is the most that can be hoped for [33, 74]. To this end we have the following theorem due to Hähner [36].

Theorem 11. *Assume $\gamma > 1$. Then D is uniquely determined by $u_\infty(\hat{x}, d)$ for $\hat{x}, d \in S^2$.*

We note that Theorem 11 remains valid if the condition on Re A is replaced by the condition

$$\overline{\xi} \cdot (\operatorname{Re} A^{-1})\xi \geq \mu |\xi|^2$$

for all $\xi \in \mathbb{C}^3$ and $x \in \overline{D}$ where μ is a positive constant such that $\mu > 1$ [7]. Note that the isotropic case when $A = I$ is handled by Theorem 10.

Uniqueness theorems for the Maxwell equations in an isotropic inhomogeneous medium have been established by Colton and Päivärinta [27] and Hähner [37]. The proof is similar to that of Theorem 10 for the scalar problem except that technical problems arise due to the fact that we must now construct a solution E, H to the Maxwell equations in an inhomogeneous isotropic medium such that E has the form

$$E(x) = e^{iz \cdot x} (\eta + r(x)),$$

where $z, \eta \in \mathbb{C}^3$, $\eta \cdot z = 0$, and $z \cdot z = k^2$. In contrast to the case of acoustic waves, it is no longer true that $r(x)$ decays to zero as $|\operatorname{Re} z|$ tends to infinity. Finally, the generalization of Theorem 11 to the case of the Maxwell equations in an anisotropic media has been done by Cakoni and Colton [5].

Historical Remarks

As previously mentioned, the first uniqueness theorem for the acoustic inverse obstacle problem was given by Schiffer in 1967 for the case of a sound-soft obstacle [61], whereas in 1988 Nachman [68], Novikov [70], and Ramm [79] established a uniqueness result for the inverse scattering problem for an inhomogeneous medium. In 1990 Isakov [40,41] proved a series of uniqueness theorems for the transmission problem with discontinuities of u across ∂D. His ideas were subsequently utilized by Kirsch, Kress, and their coworkers to establish uniqueness theorems for a variety of inverse scattering problems for both acoustic and electromagnetic waves (for references, see [22]). In particular, the proofs of Theorems 8 and 9 are based on the ideas of Kirsch and Kress [22,54].

A global uniqueness theorem for the Maxwell equations in an isotropic inhomogeneous medium was first established in 1992 by Colton and Päivärinta [27] (see also [37]). The results of [27, 37] are for the case when the magnetic permeability μ is constant. For uniqueness results in the case when μ is no longer constant, we refer to [71,72].

4 Iterative and Decomposition Methods in Inverse Scattering

Newton Iterations in Inverse Obstacle Scattering

We now turn to reconstruction methods for the inverse scattering problem for sound-soft scatterers, and as a first group we describe iterative methods. Here the inverse problem is interpreted as a nonlinear ill-posed operator equation which is solved by iteration methods such as regularized Newton methods, Landweber iterations, or conjugate gradient methods. For a fixed incident field u^i, the solution to the direct scattering problem defines the *boundary to far field operator* $\mathcal{F} : \partial D \mapsto u_\infty$ which maps the boundary ∂D of the scatterer D onto the far-field pattern u_∞ of the scattered wave u^s. In particular, $\mathcal{F}$ is the imaging operator that takes the scattering object D into its image u_∞ via the scattering process. In terms of this imaging operator, i.e., the boundary to far field operator, given a far-field pattern u_∞, the inverse problem consists in solving the operator equation

$$\mathcal{F}(\partial D) = u_\infty \tag{69}$$

for the unknown boundary ∂D. As opposed to the direct obstacle scattering problem which is linear and well-posed, the operator equation (69), i.e., the inverse obstacle scattering problem, is nonlinear and ill-posed. It is nonlinear since the solution to the direct scattering problem depends nonlinearly on the boundary, and it is ill-posed because the far-field mapping is extremely smoothing due to the analyticity of the far-field pattern.

In order to define the operator $\mathcal{F}$ properly, the most appropriate approach is to choose a fixed reference domain D and consider a family of scatterers D_h with boundaries represented in the form $\partial D_h = \{x + h(x) : x \in \partial D\}$, where $h : \partial D \to \mathbb{R}^3$ is of class C^2 and is sufficiently small in the C^2 norm on ∂D. Then we may consider the operator $\mathcal{F}$ as a mapping from a ball

$$V := \left\{h \in C^2(\partial D) : \|h\|_{C^2} < a\right\} \subset C^2(\partial D)$$

with sufficiently small radius $a > 0$ into $L^2(S^2)$. However, for ease of presentation, we proceed differently and restrict ourselves to boundaries ∂D that can be parameterized by mapping them globally onto the unit sphere S^2, i.e.,

$$\partial D = \left\{p(\hat{x}) : \hat{x} \in S^2\right\} \tag{70}$$

for some injective C^2 function $p : S^2 \to \mathbb{R}^3$. As a simple example, the reader should consider the case of starlike domains where

$$p(\hat{x}) = r(\hat{x})\hat{x}, \quad \hat{x} \in S^2, \tag{71}$$

with a radial distance function $r : S^2 \to (0, \infty)$. Then, with some appropriate subspace $W \subset C^2(S^2)$, we may interpret the operator $\mathcal{F}$ as a mapping

$$\mathcal{F} : W \to L^2(S^2), \quad \mathcal{F} : p \mapsto u_\infty,$$

and consequently the inverse obstacle scattering problem consists in solving

$$\mathcal{F}(p) = u_\infty \tag{72}$$

for the unknown function p.

Since $\mathcal{F}$ is nonlinear we may linearize

$$\mathcal{F}(p + q) = \mathcal{F}(p) + \mathcal{F}'(p)q + O\left(q^2\right)$$

in terms of a Fréchet derivative $\mathcal{F}'$. Then, given an approximation p for the solution of (72), in order to obtain an update $p + q$, we solve the approximate linear equation

$$\mathcal{F}(p) + \mathcal{F}'(p)q = u_\infty \tag{73}$$

for q. We note that the linearized equation inherits the ill-posedness of the nonlinear equation and therefore regularization is required. As in the classical Newton iterations, this linearization procedure is iterated until some stopping criteria is satisfied.

In principle the parameterization of the update $\partial D_{p+q} = \{p(\hat{x}) + q(\hat{x}) : \hat{x} \in S^2\}$ is not unique. To cope with this ambiguity, the simplest possibility is to allow only perturbations of the form

$$q(\hat{x}) = z(\hat{x})\, \nu(p(\hat{x})), \quad x \in S^2, \tag{74}$$

with a scalar function z. We denote the corresponding linear space of normal L^2 vector fields by $L^2_{\text{normal}}(S^2)$.

The Fréchet differentiability of the operator $\mathcal{F}$ is addressed in the following theorem.

Theorem 12. *The boundary to far-field mapping $\mathcal{F}: p \mapsto u_\infty$ is Fréchet differentiable. The derivative is given by*

$$\mathcal{F}'(p)q = v_{q,\infty},$$

where $v_{q,\infty}$ is the far-field pattern of the radiating solution v_q to Helmholtz equation in $\mathbb{R}^3 \setminus \overline{D}$ satisfying the Dirichlet boundary condition

$$v_q = -\nu \cdot q \, \frac{\partial u}{\partial \nu} \quad \text{on } \partial D \tag{75}$$

in terms of the total field $u = u^i + u^s$.

The boundary condition (75) for the derivative can be obtained formally by using the chain rule to differentiate the boundary condition $u = 0$ on ∂D with respect to the boundary. Extensions of Theorem 12 to the Neumann boundary condition, the perfect conductor boundary condition, and to the impedance boundary condition in acoustics and electromagnetics are also available.

To justify the application of regularization methods for stabilizing (73), injectivity and dense range of the operator $\mathcal{F}'(p) : L^2_{\text{normal}}(S^2) \to L^2(S^2)$ need to be established. This is settled for the Dirichlet condition and, for λ sufficiently large, for the impedance boundary condition and remains an open problem for the Neumann boundary condition. In the classical Tikhonov regularization, (73) is replaced by

$$\alpha q + \left[\mathcal{F}'(p)\right]^* \mathcal{F}'(p)q = \left[\mathcal{F}'(p)\right]^* \{u_\infty - \mathcal{F}(p)\} \tag{76}$$

with some positive regularization parameter α and the L^2 adjoint $[\mathcal{F}'(p)]^*$ of $\mathcal{F}'(p)$. For details on the numerical implementation, we refer to [22] and the references therein. The numerical examples strongly indicate that it is advantageous to use some Sobolev norm instead of the L^2 norm as the penalty term in the Tikhonov regularization. Numerical examples in three dimensions have been reported by Farhat et al. [31] and by Harbrecht and Hohage [38].

In closing this section on Newton iterations, we note as their main advantages that this approach is conceptually simple and, as the numerical examples in the literature indicate, leads to highly accurate reconstructions with reasonable stability against errors in the far-field pattern. On the other hand, it should be noted that for the numerical implementation, an efficient forward solver is needed and good a

priori information is required in order to ensure convergence. On the theoretical side, the convergence of regularized Newton iterations for inverse obstacle scattering problems has not been completely settled, although some progress has been made through the work of Hohage [39] and Potthast [77].

Newton-type iterations can also be employed for the simultaneous determination of the boundary shape and the impedance function λ in the impedance boundary condition (8) [58].

Decomposition Methods

The main idea of decomposition methods is to break up the inverse obstacle scattering problem into two parts: The first part deals with the ill-posedness by constructing the scattered wave u^s from its far-field pattern u_∞ and the second part deals with the nonlinearity by determining the unknown boundary ∂D of the scatterer as the location where the boundary condition for the total field $u^i + u^s$ is satisfied in a least-squares sense. In the *potential method*, for the first part, enough a priori information on the unknown scatterer D is assumed so one can place a closed surface Γ inside D. Then the scattered field u^s is sought as a single-layer potential

$$u^s(x) = \int_\Gamma \varphi(y)\Phi(x, y)\, ds(y), \quad x \in \mathbb{R}^3 \backslash \overline{D}, \tag{77}$$

with an unknown density $\varphi \in L^2(\Gamma)$. In this case the far-field pattern u_∞ has the representation

$$u_\infty(\hat{x}) = \frac{1}{4\pi} \int_\Gamma e^{-ik\,\hat{x}\cdot y} \varphi(y)\, ds(y), \quad \hat{x} \in S^2.$$

Given the far-field pattern u_∞, the density φ is now found by solving the integral equation of the first kind

$$S_\infty \varphi = u_\infty \tag{78}$$

with the compact integral operator $S_\infty : L^2(\Gamma) \to L^2(S^2)$ given by

$$(S_\infty \varphi)(\hat{x}) := \frac{1}{4\pi} \int_\Gamma e^{-ik\,\hat{x}\cdot y} \varphi(y)\, ds(y), \quad \hat{x} \in S^2.$$

Due to the analytic kernel of S_∞, the integral equation (78) is severely ill-posed. For a stable numerical solution of (78), Tikhonov regularization can be applied, i.e., the ill-posed equation (78) is replaced by

$$\alpha \varphi_\alpha + S_\infty^* S_\infty \varphi_\alpha = S_\infty^* u_\infty \tag{79}$$

with some positive regularization parameter α and the adjoint S_∞^* of S_∞.

Given an approximation of the scattered wave u_α^s obtained by inserting a solution φ_α of (79) into the potential (77), the unknown boundary ∂D is then determined by requiring the sound-soft boundary condition

$$u^i + u^s = 0 \quad \text{on } \partial D \tag{80}$$

to be satisfied in a least-squares sense, i.e., by minimizing the L^2 norm of the defect

$$\left\| u^i + u_\alpha^s \right\|_{L^2(\Lambda)}^2 \tag{81}$$

over a suitable set of admissible surfaces Λ. Instead of solving this minimization problem, one can also visualize ∂D by color coding the values of the modulus $|u|$ of the total field $u \approx u^i + u_\alpha^s$ on a sufficiently fine grid over some domain containing the scatterer.

Clearly we can expect (78) to have a solution $\varphi \in L^2(\Gamma)$ if and only if u_∞ is the far field of a radiating solution to the Helmholtz equation in the exterior of Γ with sufficiently smooth boundary values on Γ. Hence, the solvability of (78) is related to the regularity properties of the scattered wave which in general cannot be known in advance for the unknown scatterer D. Nevertheless, it is possible to provide a solid theoretical foundation to the above procedure [22, 53]. This is achieved by combining the minimization of the Tikhonov functional for (78) and the defect minimization for (81) into one cost functional

$$\|S_\infty \varphi - u_\infty\|_{L^2(S^2)}^2 + \alpha \|\varphi\|_{L^2(\Gamma)}^2 + \gamma \left\| u^i + u_\alpha^s \right\|_{L^2(\Lambda)}^2. \tag{82}$$

Here $\gamma > 0$ denotes a coupling parameter which has to be chosen appropriately for the numerical implementation in order to make the two terms in (82) be of the same magnitude, for example, $\gamma = \|u_\infty\|_{L^2(S^2)} / \|u^i\|_\infty$.

Note that the potential approach can also be employed for the inverse problem to recover the impedance given the shape of the scatterer. In this case the far-field equation (78) is solved with Γ replaced by the known boundary ∂D. After the density φ is obtained, λ can be determined in a least-squares sense from the impedance boundary condition (8) after evaluating the trace and the normal derivative of the single-layer potential (77) on ∂D.

The *point source method* of Potthast [76] can also be interpreted as a decomposition method. Its motivation is based on Huygens' principle from Theorem 4, i.e., the scattered field representation

$$u^s(x) = -\int_{\partial D} \frac{\partial u}{\partial \nu}(y)\, \Phi(x, y)\, ds(y), \quad x \in \mathbb{R}^3 \backslash \overline{D}, \tag{83}$$

and the far-field representation

$$u_\infty(\hat{x}) = -\frac{1}{4\pi} \int_{\partial D} \frac{\partial u}{\partial \nu}(y)\, e^{-ik\,\hat{x}\cdot y}\, ds(y), \quad \hat{x} \in S^2. \tag{84}$$

For $z \in \mathbb{R}^3 \setminus \overline{D}$ we choose a domain B_z such that $z \notin B_z$ and $\overline{D} \subset B_z$ and approximate the point source $\Phi(\cdot, z)$ by a *Herglotz wave function*, i.e., a superposition of plane waves such that

$$\Phi(y, z) \approx \int_{S^2} e^{ik\, y \cdot d} g_z(d)\, ds(d), \quad y \in B_z, \tag{85}$$

for some $g_z \in L^2(S^2)$. Under the assumption that there does not exist a nontrivial solution to the Helmholtz equation in B_z with homogeneous Dirichlet boundary condition on ∂B_z, the Herglotz wave functions are dense in $H^{1/2}(\partial B_z)$ [23, 30], and consequently the approximation (85) can be achieved uniformly with respect to y on compact subsets of B_z. We can now insert (85) into (32) and use (33) to obtain

$$u^s(z) \approx 4\pi \int_{S^2} g_z(\hat{x}) u_\infty(-\hat{x})\, ds(\hat{x}) \tag{86}$$

as an approximation for the scattered wave u^s. Knowing an approximation for the scattered wave, the boundary ∂D can be found as above from the boundary condition (80).

The approximation (85) can in practice be obtained by solving the ill-posed linear integral equation

$$\int_{S^2} e^{ik\, y \cdot d} g_z(d)\, ds(d) = \Phi(y, z), \quad y \in \partial B_z, \tag{87}$$

via Tikhonov regularization and the Morozov discrepancy principle. Note that although the integral equation (87) is in general not solvable, the approximation property (86) is ensured through the above denseness result on Herglotz wave functions.

An advantage of decomposition methods is that the separation of the ill-posedness and the nonlinearity is conceptually straightforward. A second and main advantage consists in the fact that their numerical implementation does not require a forward solver. As a disadvantage, as in the Newton method of the previous section, if we go beyond visualization of the level surfaces of $|u|$ and proceed with the minimization, good a priori information on the unknown scatterer is needed for the iterative solution of the optimization problem. The accuracy of the reconstructions using decomposition methods is slightly inferior to that using Newton iterations.

Iterative Methods Based on Huygens' Principle

We recall Huygens' principle (83) and (84). In view of the sound-soft boundary condition, from (83), we conclude that

$$u^i(x) = \int_{\partial D} \frac{\partial u}{\partial \nu}(y)\, \Phi(x, y)\, ds(y), \quad x \in \partial D. \tag{88}$$

Now we can interpret (84) and (88) as a system of two integral equations for the unknown boundary ∂D of the scatterer and the induced surface flux

$$\varphi := -\frac{\partial u}{\partial \nu} \quad \text{on } \partial D.$$

It is convenient to call (84) the *data equation* since it contains the given far field for the inverse problem and (88) the *field equation* since it represents the boundary condition. Both equations are linear with respect to the flux and nonlinear with respect to the boundary. Equation (84) is severely ill-posed whereas (88) is only mildly ill-posed.

Obviously there are three options for an iterative solution of (84) and (88). In a first method, given an approximation for the boundary ∂D, one can solve the mildly ill-posed integral equation of the first kind (88) for φ. Then, keeping φ fixed, Eq. (84) is linearized with respect to ∂D to update the boundary approximation. This approach has been proposed by Johansson and Sleeman [45]. In a second approach, following ideas first developed for the Laplace equation by Kress and Rundell [59], one also can solve the system (84) and (88) simultaneously for ∂D and φ by Newton iterations, i.e., by linearizing both equations with respect to both unknowns. This idea has been analyzed by Ivanyshyn and Kress [42,43]. Whereas in the first method the burden of the ill-posedness and nonlinearity is put on one equation, in a third method, a more even distribution of the difficulties is obtained by reversing the roles of (84) and (88), i.e., by solving the severely ill-posed equation (84) for φ and then linearizing (88) to obtain the boundary update. With a slight modification, this approach may also be interpreted as a decomposition method since to some extent it separates the ill-posedness and the nonlinearity. It combines the decomposition method from the previous section "Decomposition Methods" with elements of Newton iterations from section "Newton Iterations in Inverse Obstacle Scattering". Therefore, it has also been termed as a *hybrid method* and as such was analyzed by Kress and Serranho [57,85].

For a more detailed description of these three methods, using the parameterization (70), we introduce the parameterized single-layer operator and far-field operator $A, A_\infty : C^2(S^2) \times L^2(S^2) \to L^2(S^2)$ by

$$A(p,\psi)(\hat{x}) := \int_{S^2} \Phi(p(\hat{x}), p(\hat{y}))\psi(\hat{y})\, ds(\hat{y}), \quad \hat{x} \in S^2,$$

and

$$A_\infty(p,\psi)(\hat{x}) := \frac{1}{4\pi} \int_{S^2} e^{-ik\,\hat{x}\cdot p(\hat{y})}\psi(\hat{y})\, ds(\hat{y}), \quad \hat{x} \in S^2.$$

Then (84) and (88) can be written in the operator form

$$A_\infty(p,\psi) = u_\infty \tag{89}$$

and

$$A(p,\psi) = -u^i \circ p, \tag{90}$$

where we have incorporated the surface element into the density function via

$$\psi(\hat{x}) := J(\hat{x})\,\varphi(p(\hat{x})) \tag{91}$$

with the Jacobian J of the mapping p. The linearization of these equations requires the Fréchet derivatives of the operators A and A_∞ with respect to p. These can be obtained by formally differentiating their kernels with respect to p, i.e.,

$$(A'(p,\psi)q)(\hat{x}) = \int_{S^2} \operatorname{grad}_x \Phi(p(\hat{x}), p(\hat{y})) \cdot [q(\hat{x}) - q(\hat{y})]\,\psi(\hat{y})\,ds(\hat{y}), \quad x \in S^2,$$

and

$$\left(A'_\infty(p,\psi)q\right)(\hat{x}) = -\frac{ik}{4\pi} \int_{S^2} e^{-ik\,\hat{x}\cdot p(\hat{y})}\,\hat{x}\cdot q(\hat{y})\,\psi(\hat{y})\,ds(\hat{y}), \quad x \in S^2.$$

For fixed p, provided k^2 is not a Dirichlet eigenvalue of the negative Laplacian in D, both in a Hölder space setting $A(p,\cdot) : C^{0,\alpha}(S^2) \to C^{1,\alpha}(S^2)$ or in a Sobolev space setting $A(p,\cdot) : H^{-1/2}(S^2) \to H^{1/2}(S^2)$, the operator $A(p,\cdot)$ is a homeomorphism [22]. In this case, given an approximation to the boundary parameterization p, the field equation (90) can be solved for the density ψ. Then, keeping ψ fixed, linearizing the data equation (89) with respect to p leads to the linear equation

$$A'_\infty\left(p, \underbrace{[A(p,\cdot)]^{-1}(u^i \circ p)}_{-\psi}\right) q = -u_\infty - A_\infty\left(p, \underbrace{[A(p,\cdot)]^{-1}(u^i \circ p)}_{-\psi}\right) \tag{92}$$

for q to update the parameterization p via $p + q$. This procedure can be iterated.

For fixed p the operator $A'_\infty(p, [A(p,\cdot)]^{-1}(u^i \circ p))$ has a smooth kernel and therefore is severely ill-posed. This requires stabilization, for example, via Tikhonov regularization. The following theorem ensures injectivity and dense range as prerequisites for Tikhonov regularization. We recall the form (74) introduced for uniqueness of the parameterization of the update and the corresponding linear space $L^2_{\text{normal}}(S^2)$ of normal L^2 vector fields.

Theorem 13. *Assume that k^2 is not a Neumann eigenvalue of the negative Laplacian in D. Then the operator*

$$A'_\infty\left(p, [A(p,\cdot)]^{-1}(u^i \circ p)\right) : L^2_{\text{normal}}(S^2) \to L^2(S^2)$$

is injective and has dense range.

One can relate this approach to the Newton iterations for the nonlinear equation (69) for the boundary to far-field operator of section "Newton Iterations in Inverse Obstacle Scattering". In the case when k^2 is not a Dirichlet eigenvalue of the negative Laplacian in D, one can write

$$\mathcal{F}(p) = -A_\infty\left(p, [A(p,\cdot)]^{-1}(u^i \circ p)\right).$$

By the product and chain rule, this implies

$$\begin{aligned}\mathcal{F}'(p)q = &-A'_\infty\left(p, [A(p,\cdot)]^{-1}(u^i \circ p)\right) q \\ &+A_\infty\left(p, [A(p,\cdot)]^{-1} A'(p, [A(p,\cdot)]^{-1}(u^i \circ p))q\right) \\ &-A_\infty\left(p, [A(p,\cdot)]^{-1}((\operatorname{grad} u^i) \circ p)\cdot q\right).\end{aligned} \tag{93}$$

Hence, we observe a relation between the above iterative scheme and the Newton iterations for the boundary to far-field map as expressed by the following theorem.

Theorem 14. *The iteration scheme given by* (92) *can be interpreted as Newton iterations for* (69) *with the derivative of* $\mathcal{F}$ *approximated by the first term in the representation* (93).

As to be expected from this close relation to Newton iterations for (69), the quality of the reconstructions via (92) can compete with those of Newton iterations with the benefit of reduced computational costs.

The second approach for iteratively solving the system (89) and (90) consists in simultaneously linearizing both equations with respect to both unknowns. In this case, given approximations p and ψ both for the boundary parameterization and the density, the system of linear equations

$$A'_\infty(p,\psi)q + A_\infty(p,\chi) = -A_\infty(p,\psi) + u_\infty \tag{94}$$

and

$$A'(p,\psi)q + ((\operatorname{grad} u^i) \circ p)\cdot q + A(p,\chi) = -A(p,\psi) - u^i \circ p \tag{95}$$

has to be solved for q and χ in order to obtain updates $p + q$ for the boundary parameterization and $\psi + \chi$ for the density. This procedure again is iterated and coincides with Newton's method for the system (89) and (90).

For uniqueness reasons the updates must be restricted, for example, to normal fields of the form (74). Due to the smoothness of the kernels, both Eqs. (94) and (95) are severely ill-posed and require regularization with respect to both unknowns. In particular for the parameterization update, it is appropriate to incorporate penalties for Sobolev norms of q to guarantee smoothness of the boundary whereas for the density L^2 penalty terms on χ are sufficient.

The simultaneous iterations (94) and (95) again exhibit connections to the Newton iteration for (69).

Theorem 15. *Assume that k^2 is not a Dirichlet eigenvalue of the negative Laplacian in D, and set $\psi := -[A(p,\cdot)]^{-1}(u^i \circ p)$. If q satisfies the linearized boundary to far-field equation (73), then q and*

$$\chi := -[A(p,\cdot)]^{-1}(A'(p,\psi)q + ((\operatorname{grad} u^i) \circ p) \cdot q)$$

satisfy the linearized data and field equations (94) and (95). Conversely, if q and χ satisfy (94) and (95), then q satisfies (73).

Theorem 15 illustrates the difference between the iteration method based on (94) and (95) and the Newton iterations for (69). In general when performing (94) and (95) in the sequence of updates, the relation $A(p,\psi) = -(u^i \circ p)$ between the approximations p and ψ for the parameterization and the density will not be satisfied. This observation also indicates a possibility to use (94) and (95) for implementing a Newton scheme for (69). It is only necessary to replace the update $\psi + \chi$ for the density by $-[A(p+q,\cdot)]^{-1}(u^i \circ (p+q))$, i.e., at the expense of throwing away χ and solving a boundary integral equation for a new density. For a numerical implementation and three-dimensional examples, we refer to [44].

In a third method, in order to evenly distribute the burden of the ill-posedness and the nonlinearity of the inverse obstacle scattering problem, instead of solving the field equation (90) for the density and then linearizing the data equation, one can also solve the severely ill-posed data equation (91) for the density and then linearize the mildly ill-posed field equation (92) to update the boundary. In this case, given an approximation for the boundary parameterization p, first the data equation (91) is solved for the density ψ. Then, keeping ψ fixed, the field equation (92) is linearized to obtain the linear equation

$$A'(p,\psi)\, q + ((\operatorname{grad} u^i) \circ p) \cdot q = -A(p,\psi) - u^i \circ p \tag{96}$$

for q to update the parameterization p via $p + q$. This procedure of alternatingly solving (91) and (96) can be iterated. To some extent this procedure mimics a decomposition method in the sense that it decomposes the inverse problem into a severely ill-posed linear problem and a nonlinear problem.

The hybrid method suggested by Kress and Serranho [57, 85] can be considered as a slight modification of the above procedure. In this method, given an approximation p for the parameterization of the boundary, the data equation (91) is solved for the density ψ via regularization. Injectivity and dense range of the operator $A_\infty(p,\cdot) : L^2(S^2) \to L^2(S^2)$ are guaranteed provided k^2 is not a Dirichlet eigenvalue for the negative Laplacian in D [22]. Then one can define the single-layer potential

$$u^s(x) = \int_{S^2} \Phi(x, p(\hat{y}))\psi(\hat{y})\, ds(\hat{y})$$

and evaluate the boundary values of $u := u^i + u^s$ and its derivatives on the surface represented by p via the jump relations. Finally an update $p + q$ is found by linearizing the boundary condition $u \circ (p + q) = 0$, i.e., by solving the linear equation

$$u \circ p + ((\operatorname{grad} u) \circ p) \cdot q = 0 \tag{97}$$

for q. For uniqueness of the update representation, the simplest possibility is to allow only perturbations of the form (74). Then injectivity for the linear equation (97) can be established for the exact boundary.

After introducing the operator

$$\left(\tilde{A}(p,\psi)\, q\right)(\hat{x}) := \int_{S^2} \operatorname{grad}_x \Phi(p(\hat{x}), p(\hat{y})) \cdot q(\hat{x})\, \psi(\hat{y})\, ds(\hat{y}) - \frac{1}{2}\, \frac{\psi(\hat{x})\, [\nu(p(\hat{x})) \cdot q(\hat{x})]}{J(\hat{x})}$$

and observing the jump relations for the single-layer potential and (91), Eq. (97) can be rewritten as

$$\tilde{A}(p,\psi)\, q + ((\operatorname{grad} u^i) \circ p) \cdot q = -A(p,\psi) - u^i \circ p. \tag{98}$$

Comparing this with (96), we discover a relation between solving the data and field equation iteratively via (89) and (96) and the hybrid method of Kress and Serranho. In the hybrid method, the Fréchet derivative of A with respect to p is replaced by the operator $\tilde{A}$ where one linearizes only with respect to the evaluation surface for the single-layer potential but not with respect to the integration surface. For the numerical implementation of the hybrid method and numerical examples in three dimensions, we refer to [86].

All three methods of this section can be applied to the Neumann boundary condition, the perfect conductor boundary condition, and to the impedance boundary condition in acoustics and electromagnetics. They also can be employed for the simultaneous reconstruction of the boundary shape and the impedance function λ in the impedance boundary condition (8) [84].

Newton Iterations for the Inverse Medium Problem

Analogously to the inverse obstacle scattering problem, we can reformulate the inverse medium problem as a nonlinear operator equation. To this end we define the *far-field operator* $\mathcal{F} : m \mapsto u_\infty$ that maps $m := 1 - n$ to the far-field pattern u_∞ for plane wave incidence $u^i(x) = e^{ikx\cdot d}$. Since by Theorem 10 we know that m is uniquely determined by a knowledge of $u_\infty(\hat{x}, d)$ for all incident and observation directions $\hat{x}, d \in S^2$, we interpret $\mathcal{F}$ as an operator from $C(B)$ into $L^2(S^2 \times S^2)$ for a ball B that contains the unknown support of m.

In view of the Lippmann–Schwinger equation (35) and the far-field representation (36), we can write

$$(\mathcal{F}(m))(\hat{x},d) = -\frac{k^2}{4\pi}\int_B e^{-ik\,\hat{x}\cdot y} m(y)u(y,d)\,dy, \quad \hat{x},d \in S^2, \tag{99}$$

where $u(\cdot,d)$ is the unique solution of

$$u(x,d) + k^2\int_B \Phi(x,y)m(y)u(y,d)\,dy = u^i(x,d), \quad x \in B. \tag{100}$$

From (100) it can be seen that the Fréchet derivative v_q of u with respect to m (in direction q) satisfies the Lippmann–Schwinger equation

$$v_q(x,d)+k^2\int_B \Phi(x,y)[m(y)v_q(y,d)+q(y)u(y,d)]\,dy = 0, \quad x \in B. \tag{101}$$

From this and (99), it follows that the Fréchet derivative of $\mathcal{F}$ is given by

$$(\mathcal{F}'(m)\,q)(\hat{x},d) = -\frac{k^2}{4\pi}\int_B e^{-ik\,\hat{x}\cdot y}[m(y)v_q(y,d)+q(y)u(y,d)]\,dy, \quad \hat{x},d \in S^2,$$

which coincides with the far-field pattern of the solution $v_q(\cdot,d)$ of (101). Hence, we have proven the following theorem.

Theorem 16. *The far-field mapping $\mathcal{F} : m \mapsto u_\infty$ is Fréchet differentiable. The derivative is given by*

$$\mathcal{F}'(m)q = v_{q,\infty},$$

where $v_{q,\infty}$ is the far-field pattern of the radiating solution v_q to

$$\Delta v + k^2 nv = -k^2 uq \quad \text{in } \mathbb{R}^3. \tag{102}$$

This characterization of the Fréchet derivative can be used to establish injectivity of $\mathcal{F}'(m)$. We now have all the prerequisites available for a regularized Newton iteration analogous to (76).

A similar approach as that given above is also possible for the electromagnetic inverse medium problem.

Least-Squares Methods for the Inverse Medium Problem

In view of the Lippmann–Schwinger equation (35) and the far-field representation (36), the inverse medium problem is equivalent to solving the system consisting of the *field equation*

$$u(x,d) + k^2 \int_B \Phi(x,y) m(y) u(y,d)\, dy = u^i(x,d), \quad x \in B, d \in S^2, \tag{103}$$

and the *data equation*

$$-\frac{k^2}{4\pi} \int_B e^{-ik\hat{x}\cdot y} m(y) u(y,d)\, dy = u_\infty(\hat{x},d), \quad \hat{x}, d \in S^2, \tag{104}$$

where B is a ball containing the support of m. In principle one can first solve the ill-posed linear equation (104) to determine the source mu from the far-field pattern and then solve the nonlinear equation (103) to construct the contrast m. After defining the volume potential operator $T : L^2(B \times S^2) \to L^2(B \times S^2)$ and the far-field operator $F : L^2(B \times S^2) \to L^2(S^2 \times S^2)$ by

$$(Tv)(x,d) := -k^2 \int_B \Phi(x,y) v(y,d)\, dy, \quad x \in B,\ d \in S^2,$$

and

$$(Fv)(\hat{x},d) := -\frac{k^2}{4\pi} \int_B e^{-ik\hat{x}\cdot d} v(y,d)\, dy, \quad \hat{x}, d \in S^2,$$

we rewrite the field equation (103) as

$$u^i + Tmu = u \tag{105}$$

and the data equation (104) as

$$Fmu = u_\infty. \tag{106}$$

We can now define the cost function

$$\mu(m,u) := \frac{\|u^i + Tmu - u\|^2_{L^2(B\times S^2)}}{\|u^i\|^2_{L^2(B\times S^2)}} + \frac{\|u_\infty - Fmu\|^2_{L^2(S^2\times S^2)}}{\|u_\infty\|^2_{L^2(S^2\times S^2)}} \tag{107}$$

and reformulate the inverse medium problem as the optimization problem to minimize μ over the contrast $m \in V$ and the fields $u \in W$ where V and W are appropriately chosen admissible sets. The weights in the cost function are chosen such that the two terms are of the same magnitude.

This optimization problem is similar in structure to that used in (82) in connection with the decomposition method for the inverse obstacle scattering problem. However, since by Theorem 10 all incident directions are required, the discrete versions of the optimization problem suffer from a large number of unknowns. Analogous to the two step approaches of sections "Decomposition Methods" and "Iterative Methods Based on Huygens' Principle" for the inverse obstacle scattering problem,

one way to reduce the computational complexity is to treat the fields and the contrast separately, for example, by a modified conjugate gradient method as proposed by Kleinman and van den Berg [55]. In a modified version of this approach, van den Berg and Kleinman [89] transformed the Lippmann–Schwinger equation (105) into the equation

$$mu^i + mTw = w \tag{108}$$

for the contrast sources $w := mu$, and instead of simultaneously updating the contrast m and the fields u, the contrast is updated together with the contrast source w. The cost function (107) is now changed to

$$\mu(m,w) := \frac{\|mu^i + mTw - w\|^2_{L^2(B\times S^2)}}{\|u^i\|^2_{L^2(B\times S^2)}} + \frac{\|u_\infty - Fmu\|^2_{L^2(S^2\times S^2)}}{\|u_\infty\|^2_{L^2(S^2\times S^2)}}.$$

The above approach for the acoustic inverse medium problem can be adapted to the case of electromagnetic waves.

Born Approximation

The Born approximation turns the inverse medium scattering problem into a linear problem and therefore is often employed in practical applications. In view of (36), for plane wave incidence, we have the linear integral equation

$$-\frac{k^2}{4\pi}\int_{\mathbb{R}^3} e^{-ik(\hat{x}-d)\cdot y} m(y)\,dy = u_\infty(\hat{x},d) \quad \hat{x},d \in S^2. \tag{109}$$

Solving (109) for the unknown m corresponds to inverting the Fourier transform of m restricted to the ball of radius $2k$ centered at the origin, i.e., only incomplete data is available. This causes uniqueness ambiguities and leads to severe ill-posedness of the inversion. Thus, the ill-posedness which seemed to have disappeared through the inversion of the Fourier transform is back on stage. For details, we refer to [60].

A counterpart of the Born approximation in inverse obstacle scattering starts from the far field of the physical optics approximation (36) for a convex sound-soft scatterer D in the *back scattering* direction, i.e.,

$$u_\infty(-d;d) = -\frac{1}{4\pi}\int_{\nu(y)\cdot d<0} \frac{\partial}{\partial\nu(y)} e^{2ik\,d\cdot y}\,ds(y).$$

Analogously, replacing d by $-d$, we have

$$u_\infty(d;-d) = -\frac{1}{4\pi}\int_{\nu(y)\cdot d>0} \frac{\partial}{\partial\nu(y)} e^{-2ik\,d\cdot y}\,ds(y).$$

Combining the last two equations and using Green's integral theorem, we find

$$\int_{\mathbb{R}^3} \chi(y) e^{2ik\,d\cdot y}\,dy = \frac{\pi}{k^2}\left\{u_\infty(-d;d) + \overline{u_\infty(d;-d)}\right\}, \quad d \in S^2, \tag{110}$$

with the characteristic function χ of the scatterer D. Equation (110) is known as the *Bojarski identity*. Hence, in the physical optics approximation, the Fourier transform has again to be inverted from incomplete data since the physical optics approximation is valid only for large wave numbers k. For details, we refer to [60].

Historical Remarks

The boundary condition (75) was obtained by Roger [81] who first employed Newton-type iterations for the approximate solution of inverse obstacle scattering problems. Rigorous foundations for the Fréchet differentiability were given by Kirsch [47] in the sense of a domain derivative via variational methods and by Potthast [75] via boundary integral equation techniques. The potential method as a prototype of decomposition methods has been proposed by Kirsch and Kress [53]. The point source method has been suggested by Potthast [76]. The iterative methods based on Huygens' principle were introduced by Johansson and Sleeman [45], by Ivanyshyn and Kress [43] (extending a method proposed by Kress and Rundel [59] from potential theory to acoustics), and by Kress [57] and Serranho [85]. The methods described in sections "Newton Iterations for the Inverse Medium Problem," "Least Squares Methods for the Inverse Medium Problem," and "Born Approximation" have been investigated by numerous researchers over the past 30 years.

5 Qualitative Methods in Inverse Scattering

The Far-Field Operator and Its Properties

A different approach to solving inverse scattering problems than the use of iterative methods is the use of qualitative methods [7]. These methods have the advantage of requiring less a priori information than iterative methods (e.g., it is not necessary to know the topology of the scatterer or the boundary conditions satisfied by the total field) and in addition reduce a nonlinear problem to a linear problem. On the other hand, the implementation of such methods often requires more data than iterative methods do and in the case of a penetrable inhomogeneous medium only recovers the support of the scatterer together with some estimates on its material properties.

We begin by considering the scattering problem for a sound-soft obstacle (58) and (59). The *far-field operator* $F : L^2(S^2) \to L^2(S^2)$ for this problem is defined by

$$(Fg)(\hat{x}) := \int_{S^2} u_\infty(\hat{x}, d) g(d)\, ds(d), \quad \hat{x} \in S^2, \tag{111}$$

where u_∞ is the far-field pattern associated with (58) and (59). By superposition, Fg is seen to be the far-field pattern corresponding to the Herglotz wave function

$$v_g(x) := \int_{S^2} e^{ikx\cdot d} g(d)\, ds(d), \quad x \in \mathbb{R}^3, \tag{112}$$

as incident field. The function $g \in L^2(S^2)$ is known as the kernel of the Herglotz wave function. The far-field operator F is compact. It can also be shown that for the case of scattering by a sound-soft obstacle, the far-field operator is normal [7]. Of basic importance to us is the following theorem [22].

Theorem 17. *The far-field operator F corresponding to* (58) *and* (59) *is injective with dense range if and only if there does not exist a Dirichlet eigenfunction for D which is a Herglotz wave function.*

Proof. The proof is based on the reciprocity relation (61). In particular, for the L^2 adjoint $F^* : L^2(S^2) \to L^2(S^2)$, the reciprocity relation implies that

$$F^* g = \overline{RFR\bar{g}}, \tag{113}$$

where $R : L^2(S^2) \to L^2(S^2)$ is defined by $(Rg)(d) := g(-d)$. Hence, the operator F is injective if and only if its adjoint F^* is injective. Recalling that the denseness of the range of F is equivalent to the injectivity of F^*, by (113), we need only to show the injectivity of F. To this end, we note that $Fg = 0$ is equivalent to the existence of a Herglotz wave function v_g with kernel g for which the far-field pattern of the corresponding scattered field v^s is $v_\infty = 0$. By Rellich's lemma this implies that $v^s = 0$ in $\mathbb{R}^3 \backslash \overline{D}$ and the boundary condition $v_g + v^s = 0$ on ∂D now shows that $v_g = 0$ on ∂D. Since by hypothesis v_g is not a Dirichlet eigenfunction, we can conclude that $v_g = 0$ in D and hence $g = 0$. ∎

We will now turn our attention to the far-field operator associated with the inhomogeneous medium problems (65) and (68). In both cases we again define the far-field operator by (111) where u_∞ is now the far-field pattern corresponding to (65) or (68). We first consider Eq. (65) which corresponds to scattering by an inhomogeneous medium. The analogue of Theorem 17 is the following [22].

Theorem 18. *The far-field operator F corresponding to* (65) *is injective with dense range if and only if there does not exist a solution $v, w \in L^2(D), v - w \in H^2(D)$ of the* interior transmission problem

$$\Delta v + k^2 v = 0 \quad \text{in } D \tag{114}$$

$$\Delta w + k^2 n w = 0 \quad in\ D \tag{115}$$

$$v = w \quad on\ \partial D \tag{116}$$

$$\frac{\partial v}{\partial \nu} = \frac{\partial w}{\partial \nu} \quad on\ \partial D \tag{117}$$

such that v is a Herglotz wave function. Values of $k > 0$ *for which there exists a nontrivial solution of* (114)–(117) *are called* transmission eigenvalues.

A similar theorem holds for Eq. (68) which corresponds to scattering by an anisotropic medium where now (115) is replaced by

$$\nabla \cdot A \nabla w + k^2 n w = 0 \quad \text{in } D \tag{118}$$

and in (117) the normal derivative $\frac{\partial w}{\partial \nu}$ is replaced by $\nu \cdot A \nabla w$. If the coefficients in (115) or (118) are real valued, then the far-field operator is normal.

In the case of electromagnetic waves, the far-field operator becomes

$$(Fg)(\hat{x}) := \int_{S^2} E_\infty(\hat{x}, d, g(d))\, ds(d), \quad \hat{x} \in S^2, \tag{119}$$

where now $g \in L_t^2(S^2)$, the space of square integrable tangential vector fields defined on S^2, and E_∞ is the electric far-field pattern defined by (64). Theorems analogous to Theorems 17 and 18 are also valid in this case [22].

The Linear Sampling Method

The linear sampling method is a non-iterative method for solving the inverse scattering problem that was first introduced by Colton and Kirsch [19] and Colton et al. [29]. To describe this method, we first consider the case of scattering by a sound-soft obstacle, i.e., (58) and (59), and assume that for every $z \in D$, there exists a solution $g = g(\cdot, z) \in L^2(S^2)$ to the *far-field equation*

$$Fg = \Phi_\infty(\cdot, z), \tag{120}$$

where

$$\Phi_\infty(\hat{x}, z) = \frac{1}{4\pi} e^{-ik\hat{x}\cdot z}, \quad \hat{x} \in S^2.$$

Since the right-hand side of (120) is the far-field pattern of the fundamental solution (13), it follows from Rellich's lemma that

$$\int_{S^2} u^s(x, d) g(d)\, ds(d) = \Phi(x, z)$$

for $x \in \mathbb{R}^3 \setminus D$. From the boundary condition $u = 0$ on ∂D, we see that

$$v_g(x) + \Phi(x, z) = 0 \quad \text{for } x \in \partial D, \tag{121}$$

where v_g is the Herglotz wave function with kernel g. We can now conclude from (121) that v_g becomes unbounded as $z \to x \in \partial D$ and hence

$$\lim_{\substack{z \to \partial D \\ z \in D}} \|g(\cdot, z)\|_{L^2(S^2)} = \infty,$$

i.e., ∂D is characterized by points z where the solution of (120) becomes unbounded.

Unfortunately, in general, the far-field equation (120) does not have a solution nor does the above analysis say anything about what happens when $z \in \mathbb{R}^3 \setminus D$. To address these issues, we first define the single-layer operator $S : H^{-1/2}(\partial D) \to H^{1/2}(\partial D)$ by

$$(S\varphi)(x) := \int_{\partial D} \varphi(y) \Phi(x, y)\, ds(y), \quad x \in \partial D,$$

define the Herglotz operator $H : L^2(\partial D) \to H^{-1/2}(\partial D)$ as the operator mapping g to the trace of the Herglotz wave function (112) on ∂D, and let $\mathcal{F} : H^{-1/2}(\partial D) \to L^2(S^2)$ be defined by

$$(\mathcal{F}\varphi)(\hat{x}) := \int_{\partial D} \varphi(y) e^{-ik\hat{x} \cdot y}\, ds(y), \quad \hat{x} \in S^2.$$

Then, using on the one hand the fact that Herglotz wave functions are dense in the space of solutions to the Helmholtz equation in D with respect to the norm in the Sobolev space $H^1(D)$ and on the other the factorization of the far-field operator F as

$$F = -\frac{1}{4\pi} \mathcal{F} S^{-1} H,$$

one can prove the following result [7, 52].

Theorem 19. *Assume that k^2 is not a Dirichlet eigenvalue of the negative Laplacian for D, and let F be the far-field operator corresponding to (58) and (59). Then:*

1. *For $z \in D$ and a given $\epsilon > 0$, there exists $g_{z,\epsilon} \in L^2(S^2)$ such that*

$$\|F g_{z,\epsilon} - \Phi_\infty(\cdot, z)\|_{L^2(S^2)} < \epsilon$$

and the corresponding Herglotz wave function $v_{g_{z,\epsilon}}$ converges to a solution of

$$\Delta u + k^2 u = 0 \quad \text{in } D$$

$$u = -\Phi(\cdot, z) \quad \text{on } \partial D$$

in $H^1(D)$ *as* $\epsilon \to 0$.

2. *For* $z \in \mathbb{R}^3 \backslash D$ *and a given* $\epsilon > 0$, *every* $g_{z,\epsilon} \in L^2(S^2)$ *that satisfies*

$$\| F g_{z,\epsilon} - \Phi_\infty(\cdot, z) \|_{L^2(S^2)} < \epsilon$$

is such that $\lim_{\epsilon \to 0} \| v_{g_{z,\epsilon}} \|_{H^1(D)} = \infty$.

We note that the difference between cases (1) and (2) of this theorem is that for $z \in D$, the far-field pattern $\Phi_\infty(\cdot, z)$ is in the range of $\mathcal{F}$, whereas for $z \in \mathbb{R}^3 \backslash D$ this is no longer true. The *linear sampling method* is based on attempting to compute the function $g_{z,\epsilon}$ in the above theorem by using Tikhonov regularization to solve $Fg = \Phi_\infty(\cdot, z)$. In particular, one expects that the regularized solution will be relatively smaller for z in D than z in $\mathbb{R}^3 \backslash \overline{D}$, and this behavior can be visualized by color coding the values of the regularized solution on a grid over some domain containing D. A more precise statement of this observation will be made in the next section after we have discussed the factorization method for solving the inverse scattering problem. Further discussion of why linear sampling works if regularization methods are used to solve (120) can be found in [2, 3]. In addition to the inverse scattering problems (58) and (59), it is also possible to treat mixed boundary value problems as well as scattering by both isotropic and anisotropic inhomogeneous media where in the latter case we must assume that k is not a transmission eigenvalue. For full details we refer the reader to [7]. Note that in each case, it is not necessary to know the material properties of the scatterer in order to determine the support of the scatterer from a knowledge of the far-field pattern via solving the far-field equation $Fg = \Phi_\infty(\cdot, z)$.

The linear sampling method can also be extended to the case of electromagnetic waves where the far-field equation (120) is now replaced by

$$\int_{S^2} E_\infty(\hat{x}, d, g(d))\, ds(d) = E_{e,\infty}(\hat{x}, z, q),$$

where $E_\infty(\hat{x}, d, p)$ is the electric far-field pattern corresponding to the incident field (50), $g \in L_t^2(S^2)$, and $E_{e,\infty}$ is the electric far-field pattern of the electric dipole

$$E_e(x, z, q) := \frac{i}{k} \operatorname{curl}_x \operatorname{curl}_x q\Phi(x, z), \quad H_e(x, z, q) := \operatorname{curl}_x q\Phi(x, z). \tag{122}$$

Full details can be found in the lecture notes [12].

We close this section by briefly describing a version of the linear sampling method based on the reciprocity gap functional which is applicable to objects situated in a piecewise homogeneous background medium. Assume that an unknown

scattering object is embedded in a portion B of a piecewise inhomogeneous medium where the index of refraction is constant with wave number k. Let $B_0 \subset B$ be a domain in B having a smooth boundary ∂B_0 such that the scattering obstacle D satisfies $D \subset B_0$, and let ν be the unit outward normal to ∂B_0. We now define the *reciprocity gap functional* by

$$R(u, v) := \int_{\partial B_0} \left(u \frac{\partial v}{\partial \nu} - v \frac{\partial u}{\partial \nu} \right) ds,$$

where u and v are solutions of the Helmholtz equation in $B_0 \backslash \overline{D}$ and $u, v \in C^1(\overline{B_0 \backslash D})$. In particular, we want u to be the total field due to a point source situated at $x_0 \in B \backslash \overline{B_0}$ and $v = v_g$ to be a Herglotz wave function with kernel g. We then consider the integral equation

$$R(u, v_g) = R(u, \Phi_z),$$

where $\Phi_z := \Phi(\cdot, z)$ is the fundamental solution (13) and $u = u(\cdot, x_0)$ where x_0 is now assumed to be on a smooth surface C in $B \backslash \overline{B_0}$ that is homotopic to ∂B_0. If D is a sound-soft obstacle, we assume that k^2 is not a Dirichlet eigenvalue of the negative Laplacian in D, and if D is an isotropic inhomogeneous medium, we assume that k is not a transmission eigenvalue. We then have the following theorem [17].

Theorem 20. *Assume that the above assumptions on D are satisfied. Then:*

1. *If $z \in D$, then there exists a sequence $\{g_n\}$ in $L^2(S^2)$ such that*

$$\lim_{n \to \infty} R(u, v_{g_n}) = R(u, \Phi_z), \quad x_0 \in C,$$

   and v_{g_n} converges in $L^2(D)$.
2. *If $z \in B_0 \backslash D$, then for every sequence $\{g_n\}$ in $L^2(S^2)$ such that*

$$\lim_{n \to \infty} R(u, v_{g_n}) = R(u, \Phi_z), \quad x_0 \in C,$$

   we have that $\lim_{n \to \infty} \|v_{g_n}\|_{L^2(D)} = \infty$.

In particular, Theorem 20 provides a method for determining D from a knowledge of the Cauchy data of u on ∂B_0 in a manner analogous to that of the linear sampling method. Numerical examples using this method can be found in [17]. The extension of Theorem 20 to the Maxwell equations, together with numerical examples, can be found in [13].

The Factorization Method

The linear sampling method is complicated by the fact that in general $\Phi_\infty(\cdot,z)$ is not in the range of the far-field operator F for either $z \in D$ or $z \in \mathbb{R}^3\backslash\overline{D}$. For the case of acoustic waves when F is normal (e.g., the scattering problem corresponding to (58) and (59) or (65) for n real valued), the problem was resolved by Kirsch in [48, 49] who proposed replacing the far field equation $Fg = \Phi_\infty(\cdot,z)$ by

$$(F^*F)^{1/4}g = \Phi_\infty(\cdot,z), \tag{123}$$

where F^* is again the adjoint of F in $L^2(S^2)$. In particular, if $G : H^{1/2}(\partial D) \to L^2(S^2)$ is defined by $Gf = v_\infty$ where v_∞ is the far field pattern of the solution to the radiating exterior Dirichlet problem (see Theorem 3) with boundary data $f \in L^2(\partial D)$, then the following theorem is valid [48].

Theorem 21. *Assume that k^2 is not a Dirichlet eigenvalue of the negative Laplacian for D. Then the ranges of $G : H^{1/2}(\partial D) \to L^2(S^2)$ and $(F^*F)^{1/4} : L^2(S^2) \to L^2(S^2)$ coincide.*

A result analogous to Theorem 21 is also valid for the scattering problem corresponding to (65) for n real valued where we now must assume the k is not an interior transmission eigenvalue [49]. Note that Theorem 21 provides an alternate method to the linear sampling method for solving the inverse scattering problem corresponding to the scattering of acoustic waves by a sound-soft obstacle. This follows from the fact that $\Phi_\infty(\cdot,z)$ is in the range of G if and only if $z \in D$, i.e., Eq. (123) is solvable if and only if $z \in D$. This is an advantage over the linear sampling method since if (123) is solved by using Tikhonov regularization, then as the noise level on u_∞ tends to zero, the norm of the regularized solution remains bounded if and only if $z \in D$. A similar statement cannot be made if regularization methods are used to solve $Fg = \Phi_\infty(\cdot,z)$. However, using Theorem 21, the following theorem has been established by Arens and Lechleiter [3] (see also [52]).

Theorem 22. *Let F be the far-field operator associated with the scattering problems (58) and (59), and assume that k^2 is not a Dirichlet eigenvalue of the negative Laplacian for D. For $z \in D$, let $g_z \in L^2(S^2)$ be the solution of $(F^*F)^{1/4}g_z = \Phi_\infty(\cdot,z)$, and for every $z \in \mathbb{R}^3$ and $\epsilon > 0$, let $g_{z,\epsilon}$ be the solution of $Fg = \Phi_\infty(\cdot,z)$ obtained by Tikhonov regularization, i.e., the unique solution of $\epsilon g + F^*Fg = F^*\Phi_\infty$. Then the following statements are valid:*

1. *Let $v_{g_{z,\epsilon}}$ be the Herglotz wave function with kernel $g_{z,\epsilon}$. Then for every $z \in D$, the limit $\lim_{\epsilon\to 0} v_{g_{z,\epsilon}}(z)$ exists. Furthermore, there exists $c > 0$, depending only on F, such that for every $z \in D$ we have that*

$$c\|g_z\|^2_{L^2(S^2)}\| \le \lim_{\epsilon\to 0} \left|v_{g_{z,\epsilon}}(z)\right| \le \|g_z\|^2_{L^2(S^2)}.$$

2. *For $z \notin D$ we have that $\lim_{\epsilon\to 0} v_{g_{z,\epsilon}}(z) = \infty$.*

Using Theorem 21 to solve the inverse scattering problem associated with the scattering problems (58) and (59) is called the *factorization method*. This method has been extended to a wide variety of scattering problems for both acoustic and electromagnetic waves, and for details we refer the reader to [52]. Since this method and its generalizations are fully discussed in the chapter in this handbook on sampling methods, we will not pursue the topic further here. A drawback of both the linear sampling method and the factorization method is the large amount of data needed for the inversion procedure. In particular, although the linear sampling method can be applied for limited aperture far-field data, one still needs multistatic data defined on an open subset of S^2.

Lower Bounds for the Surface Impedance

One of the advantages that the linear sampling method has over other qualitative methods in inverse scattering theory is that the far-field equation can not only be used to determine the support of the scatterer but in some circumstances can also be used to obtain lower bounds on the constitutive parameters of the scattering object. In this section we will consider two such problems: the determination of the surface impedance of a partially coated object and the determination of the index of refraction of a non-absorbing scatterer. In the first case, we will need to consider a mixed boundary value problem for the Helmholtz equation, whereas in the second case, we will need to investigate the spectral properties of the interior transmission problem introduced in Theorem 18 of the previous section.

Mixed boundary value problems typically model the scattering by objects that are coated by a thin layer of material on part of the boundary. In the study of inverse problems for partially coated obstacles, it is important to mention that, in general, it is not known a priori whether or not the scattering object is coated and if so what the extent of the coating is. We will focus our attention in this section on the special case when on the coated part of the boundary, the total field satisfies an impedance boundary condition and on the remaining part of the boundary, the total field (or the tangential component in the case of electromagnetic waves) vanishes. This corresponds to the case when a perfect conductor is partially coated by a thin dielectric layer. For other mixed boundary value problems in scattering theory and their associated inverse problems, we refer the reader to [7] and the references contained therein.

Let $D \subset \mathbb{R}^3$ be as described in Sect. 1, and let ∂D be dissected as $\partial D = \Gamma_D \cup \Pi \cup \Gamma_I$ where Γ_D and Γ_I are disjoint, relatively open subsets of ∂D having Π as their common boundary. Let $\lambda \in L_\infty(\Gamma_I)$ be such that $\lambda(x) \geq \lambda_0 > 0$ for all $x \in \Gamma_I$. We consider the scattering problem for the Helmholtz equation (58) where $u = u^i + u^s$ satisfies the boundary condition

$$\begin{aligned} u &= 0 \quad \text{on } \Gamma_D, \\ \frac{\partial u}{\partial \nu} + i\lambda u &= 0 \quad \text{on } \Gamma_I, \end{aligned} \tag{124}$$

$u^i(x) = e^{ikx\cdot d}$, and u^s is a radiating solution. It can be shown that this direct scattering problem has a unique solution in $H_{\rm loc}(\mathbb{R}^3\setminus\overline{D})$ [7]. We again define the far-field operator by (111) where u_∞ is now the far-field pattern corresponding to the boundary condition (124).

In [7] it is shown that there exists a unique solution $u_z \in H^1(D)$ of the interior mixed boundary value problem

$$\Delta u_z + k^2 u_z = 0 \quad \text{in } D \tag{125}$$

$$u_z + \Phi(\cdot, z) = 0 \quad \text{on } \Gamma_D \tag{126}$$

$$\frac{\partial}{\partial \nu}\left(u_z + \Phi(\cdot, z)\right) + i\lambda\left(u_z + \Phi(\cdot, z)\right) = 0 \quad \text{on } \Gamma_I \tag{127}$$

for $z \in D$ where Φ is the fundamental solution to the Helmholtz equation. Then, if $\Phi_\infty(\cdot, z)$ is the far-field pattern of $\Phi(\cdot, z)$, we have the following theorem [7].

Theorem 23. *Let $\epsilon > 0$, $z \in D$, and u_z be the unique solution of (125–127). Then there exists a Herglotz wave function $v_{g_{z,\epsilon}}$ with kernel $g_{z,\epsilon} \in L^2(S^2)$ such that*

$$\|u_z - v_{g_{z,\epsilon}}\|_{H^1(D)} \le \epsilon.$$

Moreover, there exists a positive constant c independent of ϵ such that

$$\|Fg_{z,\epsilon} - \Phi_\infty(\cdot, z)\|_{L^2(S^2)} \le c\epsilon.$$

We can now use Green's formula to show that [6]

$$\int_{\partial D} \lambda |u_z + \Phi(\cdot, z)|^2 \, ds = -\frac{k}{4\pi} - \operatorname{Im} u_z(z).$$

From this we immediately deduce the inequality

$$\|\lambda\|_{L^\infty(\Gamma_I)} \ge \frac{-k/4\pi - \operatorname{Im} u_z(z)}{\|u_z + \Phi(\cdot, z)\|^2_{L^2(\partial D)}}. \tag{128}$$

How is the inequality (128) of practical use? To evaluate the right-hand side of (128), we need to know ∂D and u_z. Both are determined by solving the far field equation $Fg = \Phi_\infty(\cdot, z)$ using Tikhonov regularization and then using the linear sampling method to determine ∂D and the regularized solution $g \in L^2(S^2)$ to construct the Herglotz wave function v_g. By Theorem 23 we expect that v_g is an approximation to u_z. However, at this time, there is no analogue of Theorem 22 for the mixed boundary value problem, and hence this is not guaranteed. Nevertheless in all numerical experiments to date, this approximation appears to be remarkably accurate and thus allows us to obtain a lower bound for $\|\lambda\|_{L^\infty(\Gamma_I)}$ via (128).

The corresponding scattering problem for the Maxwell equations is to find a solution $E = E^i + E^s$ to (62) satisfying the mixed boundary condition

$$\begin{aligned} \nu \times E &= 0 \quad \text{on } \Gamma_D \\ \nu \times \operatorname{curl} E - i\lambda(\nu \times E) \times \nu &= 0 \quad \text{on } \Gamma_I, \end{aligned} \tag{129}$$

where E^i is the plane wave (50) and E^s is radiating. The existence of a unique solution E in an appropriate Sobolev space is shown in [11]. We again define the far-field operator by (119) where E_∞ is now the electric far-field pattern corresponding to (129). Analogous to (125–127) we now have the interior mixed boundary value problem

$$\operatorname{curl}\operatorname{curl} E_z - k^2 E_z = 0 \quad \text{in } D \tag{130}$$

$$\nu \times [E_z + E_e(\cdot, z, q)] = 0 \quad \text{on } \Gamma_D \tag{131}$$

$$\nu \times \operatorname{curl}[E_z + E_e(\cdot, z, q)] - i\lambda\,[\nu \times (E_z + E_e(\cdot, z, q))] = 0 \quad \text{on } \Gamma_I, \tag{132}$$

where $z \in D$ and E_e is the electric dipole defined by (122). The existence of a unique solution to (130–132) in an appropriate Sobolev space is established in [11]. From the analysis in [6], we have the inequality

$$\|\lambda\|_{L^\infty(\Gamma_I)} \geq \frac{-k^2|q|^2/6\pi + k\,\operatorname{Re}(q \cdot E_z)}{\|E_z + E_e(\cdot, z, q)\|^2_{L^2_t(\partial D)}} \tag{133}$$

analogous to (128) for the Helmholtz equation. For numerical examples using (133), we refer the reader to [26].

Similar inequalities as those derived above for the impedance boundary value problem can also be obtained for the *conductive boundary value problem*, i.e., the case when a dielectric is partially coated by a thin, highly conducting layer [7, 26].

Transmission Eigenvalues

We have previously encountered transmission eigenvalues in Theorem 18 where they were connected with the injectivity and dense range of the far-field operator. In this section we shall examine transmission eigenvalues and the interior transmission problem in more detail. This investigation is particularly relevant to the inverse scattering problem since transmission eigenvalues can be determined from the far field pattern [10] and, as will be seen, can be used to obtain lower bounds for the index of refraction.

We begin by considering the interior transmission problem (114–117) from Theorem 18 and will be concerned with the existence and countability of transmission eigenvalues. The existence of transmission eigenvalues was first established by Päivärinta and Sylvester [73], and their results were strengthened by Cakoni et al. [14].

Theorem 24. *Assume that n is real valued such that $n(x) > 1$ for all $x \in \overline{D}$ or $0 < n(x) < 1$ for all $x \in \overline{D}$. Then there exist an infinite number of transmission eigenvalues.*

We note that it can be shown that as $\sup_{x\in D} |n(x) - 1| \to 0$, then the first transmission eigenvalue tends to infinity, i.e., in the Born approximation transmission, eigenvalues do not exist [28].

Similar results as in Theorem 24 can be obtained for an anisotropic medium and for the Maxwell equations [14].

By Theorem 24 the existence of transmission eigenvalues is established. It can also be shown that the set of transmission eigenvalues is discrete [15, 20, 50, 83]. The following theorem [28] establishes a lower bound for the first transmission eigenvalue which is reminiscent of the famous *Faber–Krahn inequality* for the first Dirichlet eigenvalue for the negative Laplacian (which we denote by λ_1).

Theorem 25. *Assume that $n(x) > 1$ for $x \in \overline{D}$, and let $k_1 > 0$ be the first transmission eigenvalue for the interior transmission problem* (114–117)*. Then*

$$k_1^2 \geq \frac{\lambda_1(D)}{\sup\limits_{x\in D} n(x)}.$$

Theorem 25 has been generalized to the case of anisotropic media and the Maxwell equations [8].

Finally, in the case of the interior transmission problem (114–117) where there are cavities in D, i.e., regions $D_0 \subset D$ where $n(x) = 1$ for $x \in D_0$, it can be shown that transmission eigenvalues exist and form a discrete set and the first transmission eigenvalue k_1 satisfies [9]

$$k_1^2 \geq \frac{\lambda_1(D)}{\sup\limits_{x\in D\setminus D_0} n(x)}.$$

Note that since in each of the above cases D can be determined by the linear sampling method, $\lambda_1(D)$ is known, and hence given k_1, the above inequalities yield a lower bound for the supremum of the index of refraction.

Historical Remarks

The use of qualitative methods to solve inverse scattering problems began with the 1996 paper of Colton and Kirsch [19] and the 1997 paper of Colton et al. [29]. These papers were in turn motivated by the dual space method of Colton and Monk developed in [24, 25]. Both [19] and [29] were concerned with the case of scattering of acoustic waves. The extension of the linear sampling method to electromagnetic waves was first outlined by Kress [56] and then discussed in more detail by Colton

et al. [18] and Haddar and Monk [34]. The factorization method was introduced in 1998 and 1999 by Kirsch [48, 49] for acoustic scattering problems. Attempts to extend the factorization method to the case of electromagnetic waves have been only partly successful. In particular, the factorization method for the scattering of electromagnetic waves by a perfect conductor remains an open question.

In addition to the linear sampling and factorization methods, there have been a number of other qualitative methods developed primarily by Ikehata and Potthast and their coworkers. Although space is too short to discuss these alternate qualitative methods in this survey, we refer the reader to [52, 78] for details and references.

The countability of transmission eigenvalues for acoustic waves was established by Colton et al. [20] and Rynne and Sleeman [83] and for the Maxwell equations by Cakoni and Haddar [15] and Kirsch [50]. The existence of transmission eigenvalues for acoustic waves was first given by Päivärinta and Sylvester [73] for the isotropic case and for the anisotropic case by Cakoni and Haddar [16] and Kirsch [51] who also established the existence of transmission eigenvalues for Maxwell's equations. These results were subsequently improved by Cakoni et al. [14]. Inequalities for the first transmission eigenvalues were first obtained by Colton et al. [28] and Cakoni et al. [8, 9].

Cross-References

- ▸ Electrical Impedance Tomography
- ▸ EM Algorithms
- ▸ Iterative Solution Methods
- ▸ Synthetic Aperture Radar Imaging
- ▸ Regularization Methods for Ill-Posed Problems
- ▸ Sampling Methods
- ▸ Tomography
- ▸ Wave Phenomena

References

1. Alessandrini, G., Rondi, L.: Determining a sound–soft polyhedral scatterer by a single far–field measurement. Proc. Am. Math. Soc. **133**, 1685–1691 (2005)
2. Arens, T.: Why linear sampling works. Inverse Prob. **20**, 163–173 (2004)
3. Arens, T., Lechleiter, A.: The linear sampling method revisited. J. Integral Eqn. Appl. **21**, 179–202 (2009)
4. Bukhgeim, A.: Recovering a potential from Cauchy data in the two-dimensional case. J. Inverse Ill–Posed Prob. **16**, 19–33 (2008)
5. Cakoni, F., Colton, D.: A uniqueness theorem for an inverse electromagnetic scattering problem in inhomogeneous anisotropic media. Proc. Edinb. Math. Soc. **46**, 293–314 (2003)
6. Cakoni, F., Colton, D.: The determination of the surface impedance of a partially coated obstacle from far field data. SIAM J. Appl. Math. **64**, 709–723 (2004)
7. Cakoni, F., Colton, D.: Qualitative Methods in Inverse Scattering Theory. Springer, Berlin (2006)

8. Cakoni, F., Colton, D., Haddar, H.: The computation of lower bounds for the norm of the index of refraction in anisotropic media from far field data. J. Integral Eqn. Appl. **21**, 203–227 (2009)
9. Cakoni, F., Colton, D., Haddar, H.: The interior transmission problem for regions with cavities. SIAM J. Math. Anal. **42**, 145–162 (2010)
10. Cakoni, F., Colton, D., Haddar, H.: On the determination of Dirichlet and transmission eigenvalues from far field data. Comput. Rend. Math. **348**, 379–383 (2010)
11. Cakoni, F., Colton, D., Monk, P.: The electromagnetic inverse scattering problem for partly coated Lipschitz domains. Proc. R. Soc. Edinb. **134A**, 661–682 (2004)
12. Cakoni, F., Colton, D., Monk, P.: The Linear Sampling Method in Inverse Electromagnetic Scattering. SIAM.
13. Cakoni, F., Fares, M., Haddar, H.: Analysis of two linear sampling methods applied to electromagnetic imaging of buried objects. Inverse Prob. **22**, 845–867 (2006)
14. Cakoni, F., Gintides, D., Haddar, H.: The existence of an infinite discrete set of transmission eigenvalues. SIAM J. Math. Anal. **42**, 237–255 (2010)
15. Cakoni, F., Haddar, H.: A variational approach for the solution of the electro-magnetic interior transmission problem for anisotropic media. Inverse Prob. Imaging **1**, 443–456 (2007)
16. Cakoni, F., Haddar, H.: On the existence of transmission eigenvalues in an inhomogeneous medium. Appl. Anal. **89**, 29–47 (2010)
17. Colton, D., Haddar, H.: An application of the reciprocity gap functional to inverse scattering theory. Inverse Prob. **21**, 383–398 (2005)
18. Colton, D., Haddar, H., Monk, P.: The linear sampling method for solving the electromagnetic inverse scattering problem. SIAM J. Sci. Comput. **24**, 719–731 (2002)
19. Colton, D., Kirsch, A.: A simple method for solving inverse scattering problems in the resonance region. Inverse Prob. **12**, 383–393 (1996)
20. Colton, D., Kirsch, A., Päivärinta, L.: Far field patterns for acoustic waves in an inhomogeneous medium. SIAM J. Math. Anal. **20**, 1472–1483 (1989)
21. Colton, D., Kress, R.: Eigenvalues of the far field operator for the Helmholtz equation in an absorbing medium. SIAM J. Appl. Math. **55**, 1724–1735 (1995)
22. Colton, D., Kress, R.: Inverse Acoustic and Electromagnetic Scattering Theory, 2nd edn. Springer, Berlin (1998)
23. Colton, D., Kress, R.: On the denseness of Herglotz wave functions and electromagnetic Herglotz pairs in Sobolev spaces. Math. Methods Appl. Sci. **24**, 1289–1303 (2001)
24. Colton, D., Monk, P.: A novel method for solving the inverse scattering problem for time harmonic acoustic waves in the resonance region II. SIAM J. Appl. Math. **26**, 506–523 (1986)
25. Colton, D., Monk, P.: The inverse scattering problem for acoustic waves in an inhomogeneous medium. Quart. J. Mech. Appl. Math. **41**, 97–125 (1988)
26. Colton, D., Monk, P.: Target identification of coated objects. IEEE Trans. Antennas Prop. **54**, 1232–1242 (2006)
27. Colton, D., Päivärinta, L.: The uniqueness of a solution to an inverse scattering problem for electromagnetic waves. Arch. Ration. Mech. Anal. **119**, 59–70 (1992)
28. Colton, D., Päivärinta, L., Sylvester, J.: The interior transmission problem. Inverse Probl. Imaging **1**, 13–28 (2007)
29. Colton, D., Piana, M., Potthast, R.: A simple method using Mozorov's discrepancy principle for solving inverse scattering problems. Inverse Prob. **13**, 1477–1493 (1997)
30. Colton, D., Sleeman, B.: An approximation property of importance in inverse scattering theory. Proc. Edinburgh. Math. Soc. **44**, 449–454 (2001)
31. Farhat, C., Tezaur, R., Djellouli, R.: On the solution of three-dimensional inverse obstacle acoustic scattering problems by a regularized Newton method. Inverse Prob. **18**, 1229–1246 (2002)
32. Gintides, D.: Local uniqueness for the inverse scattering problem in acoustics via the Faber–Krahn inequality. Inverse Prob. **21**, 1195–1205 (2005)
33. Gylys–Colwell, F.: An inverse problem for the Helmholtz equation. Inverse Prob. **12**, 139–156 (1996)

34. Haddar, H., Monk, P.: The linear sampling method for solving the electromagnetic inverse medium problem. Inverse Prob. **18**, 891–906 (2002)
35. Hähner, P.: A periodic Faddeev–type solution operator. J. Diff. Eqn. **128**, 300–308 (1996)
36. Hähner, P.: On the uniqueness of the shape of a penetrable anisotropic obstacle. J. Comput. Appl. Math. **116**, 167–180 (2000)
37. Hähner, P.: Electromagnetic wave scattering. In: Pike, R., Sabatier, P. (eds.) Scattering. Academic, New York (2002)
38. Harbrecht, H., Hohage, T.: Fast methods for three-dimensional inverse obstacle scattering problems. J. Integral Eqn. Appl. **19**, 237–260 (2007)
39. Hohage, T.: Iterative methods in inverse obstacle scattering: regularization theory of linear and nonlinear exponentially ill-posed problems. Dissertation, Linz (1999)
40. Isakov, V.: On the uniqueness in the inverse transmission scattering problem. Commun. Partial Diff. Eqns. **15**, 1565–1587 (1988)
41. Isakov, V.: Inverse Problems for Partial Differential Equations. Springer, Berlin (1996)
42. Ivanyshyn, O.: Nonlinear boundary integral equations in inverse scattering. Dissertation, Gäottingen (2007)
43. Ivanyshyn, O., Kress, R.: Nonlinear integral equations in inverse obstacle scattering. In: Fotiatis M (ed) Mathematical Methods in Scattering Theory and Biomedical Engineering. World Scientific, Singapore, pp. 39–50 (2006)
44. Ivanyshyn, O., Kress, R.: Identification of sound-soft 3D obstacles from phaseless data. Inverse Prob. Imaging **4**, 131–149 (2010)
45. Johansson, T., Sleeman, B.: Reconstruction of an acoustically sound-soft obstacle from one incident field and the far field pattern. IMA J. Appl. Math. **72**, 96–112 (2007)
46. Jones, D.S.: Acoustic and Electromagnetic Waves. Clarendon, Oxford (1986)
47. Kirsch, A.: The domain derivative and two applications in inverse scattering. Inverse Prob. **9**, 81–86 (1993)
48. Kirsch, A.: Characterization of the shape of a scattering obstacle using the spectral data of the far field operator. Inverse Prob. **14**, 1489–1512 (1998)
49. Kirsch, A.: Factorization of the far field operator for the inhomogeneous medium case and an application in inverse scattering theory. Inverse Prob. **15**, 413–429 (1999)
50. Kirsch, A.: An integral equation approach and the interior transmission problem for Maxwell's equations. Inverse Prob. Imaging **1**, 159–179 (2007)
51. Kirsch, A.: On the existence of transmission eigenvalues. Inverse Prob. Imaging **3**, 155–172 (2009)
52. Kirsch, A., Grinberg, N.: The Factorization Method for Inverse Problems. Oxford University Press, Oxford (2008)
53. Kirsch, A., Kress, R.: An optimization method in inverse acoustic scattering. In: Brebbia, C.A. et al. (eds.) Boundary Elements IX. Fluid Flow and Potential Applications, vol. 3. Springer, Berlin (1987)
54. Kirsch, A., Kress, R.: Uniqueness in inverse obstacle scattering. Inverse Prob. **9**, 285–299 (1993)
55. Kleinman, R., van den Berg, P.: A modified gradient method for two dimensional problems in tomography. J. Comput. Appl. Math. **42**, 17–35 (1992)
56. Kress, R.: Electromagnetic waves scattering. In: Pike, R., Sabatier, P. (eds.) Scattering. Academic, New York (2002)
57. Kress, R.: Newton's Method for inverse obstacle scattering meets the method of least squares. Inverse Prob. **19**, 91–104 (2003)
58. Kress, R., Rundell, W.: Inverse scattering for shape and impedance. Inverse Prob. **17**, 1075–1085 (2001)
59. Kress, R., Rundell, W.: Nonlinear integral equations and the iterative solution for an inverse boundary value problem. Inverse Prob. **21**, 1207–1223 (2005)
60. Langenberg, K.: Applied inverse problems for acoustic, electromagnetic and elastic wave scattering. In: Sabatier, P. (ed.) Basic Methods of Tomography and Inverse Problems. Adam Hilger, Bristol (1987)

61. Lax, P.D., Phillips, R.S.: Scattering Theory. Academic, New York (1967)
62. Liu, H.: A global uniqueness for formally determined electromagnetic obstacle scattering. Inverse Prob. **24**, 035018 (2008)
63. Liu, H., Zou, J.: Uniqueness in an inverse acoustic obstacle scattering problem for both sound-hard and sound-soft polyhedral scatterers. Inverse Prob. **22**, 515–524 (2006)
64. McLean, W.: Strongly Elliptic Systems and Boundary Integral Equations. Cambridge University Press, Cambridge (2000)
65. Monk, P.: Finite Element Methods for Maxwell's Equations. Oxford University Press, Oxford (2003)
66. Morse, P.M., Ingard, K.U.: Linear acoustic theory. In: Faugge, S. (ed.) Encyclopedia of Physics. Springer, Berlin (1961)
67. Mäuller, C.: Foundations of the Mathematical Theory of Electromagnetic Waves. Springer, Berlin (1969)
68. Nachman, A.: Reconstructions from boundary measurements. Ann. Math. **128**, 531–576 (1988)
69. Nédélec, J.C.: Acoustic and Electromagnetic Equations. Springer, Berlin (2001)
70. Novikov, R.: Multidimensional inverse spectral problems for the equation $-\Delta\psi + (v(x) - Eu(x))\,\psi = 0$. Trans. Funct. Anal. Appl. **22**, 263–272 (1988)
71. Ola, P., Päivärinta, L., Somersalo, E.: An inverse boundary value problem in electrodynamics. Duke Math. J. **70**, 617–653 (1993)
72. Ola, P., Somersalo, E.: Electromagnetic inverse problems and generalized Sommerfeld potentials. SIAM J. Appl. Math. **56**, 1129–1145 (1996)
73. Päivärinta, L., Sylvester, J.: Transmission eigenvalues. SIAM J. Math. Anal. **40**, 738–753 (2008)
74. Piana, M.: On uniqueness for anisotropic inhomogeneous inverse scattering problems. Inverse Prob. **14**, 1565–1579 (1998)
75. Potthast, R.: Fréchet differentiability of boundary integral operators in inverse acoustic scattering. Inverse Prob. **10**, 431–447 (1994)
76. Potthast, R.: Point-Sources and Multipoles in Inverse Scattering Theory. Chapman and Hall, London (2001)
77. Potthast, R.: On the convergence of a new Newton-type method in inverse scattering. Inverse Prob. **17**, 1419–1434 (2001)
78. Potthast, R.: A survey on sampling and probe methods for inverse problems. Inverse Prob. **22**, R1–R47 (2006)
79. Ramm, A.: Recovery of the potential from fixed energy scattering data. Inverse Prob. **4**, 877–886 (1988)
80. Rjasanow, S., Steinbach, O.: The Fast Solution of Boundary Integral Equations. Springer, Berlin (2007)
81. Roger, R.: Newton Kantorovich algorithm applied to an electromagnetic inverse problem. IEEE Trans. Antennas Prop. **29**, 232–238 (1981)
82. Rondi, L.: Unique determination of non-smooth sound-soft scatterers by finitely many far-field measurements. Indiana Univ. Math. J. **52**, 1631–1662 (2003)
83. Rynne, B.P., Sleeman, B.D.: The interior transmission problem and inverse scattering from inhomogeneous media. SIAM J. Math. Anal. **22**, 1755–1762 (1991)
84. Serranho, P.: A hybrid method for inverse scattering for shape and impedance. Inverse Prob. **22**, 663–680 (2006)
85. Serranho, P.: A hybrid method for inverse obstacle scattering problems. Dissertation, Gäottingen (2007)
86. Serranho, P.: A hybrid method for sound-soft obstacles in 3D. Inverse Prob. Imaging **1**, 691–712 (2007)
87. Stefanov, P., Uhlmann, G.: Local uniqueness for the fixed energy fixed angle inverse problem in obstacle scattering. Proc. Am. Math. Soc. **132**, 1351–1354 (2003)
88. Sylvester, J., Uhlmann, G.: A global uniqueness theorem for an inverse boundary value problem. Ann. Math. **125**, 153–169 (1987)
89. van den Berg, R., Kleinman, R.: A contrast source inversion method. Inverse Prob. **13**, 1607–1620 (1997)

Electrical Impedance Tomography

Andy Adler, Romina Gaburro, and William Lionheart

Contents

1 Introduction 702
Measurement Systems and Physical Derivation 704
The Concentric Anomaly: A Simple Example 708
Measurements with Electrodes 710
2 Uniqueness and Stability of the Solution 713
The Isotropic Case 714
The Anisotropic Case 731
Some Remarks on the Dirichlet-to-Neumann Map 739
3 The Reconstruction Problem 741
Locating Objects and Boundaries 741
Forward Solution 744
Regularized Linear Methods 747
Regularized Iterative Nonlinear Methods 749
Direct Nonlinear Solution 754
4 Conclusion 757
References 758

Abstract

This chapter reviews the state of the art and the current open problems in electrical impedance tomography (EIT), which seeks to recover the conductivity

A. Adler (✉)
Systems and Computer Engineering, Clarkson University, Ottawa, ON, Canada
e-mail: adler@sce.carleton.ca

R. Gaburro
University of Limerick, Limerick, Ireland
e-mail: romina.gaburro@ul.ie

W. Lionheart
The University of Manchester, Manchester, UK
e-mail: bill.lionheart@manchester.ac.uk

© Springer Science+Business Media New York 2015

O. Scherzer (ed.), *Handbook of Mathematical Methods in Imaging*,
DOI 10.1007/978-1-4939-0790-8_14

(or conductivity and permittivity) of the interior of a body from knowledge of electrical stimulation and measurements on its surface. This problem is also known as the inverse conductivity problem and its mathematical formulation is due to A. P. Calderón, who wrote in 1980, the first mathematical formulation of the problem, "On an inverse boundary value problem." EIT has interesting applications in fields such as medical imaging (to detect air and fluid flows in the heart and lungs and imaging of the breast and brain) and geophysics (detection of conductive mineral ores and the presence of ground water). It is well known that this problem is severely ill-posed, and thus this chapter is devoted to the study of the uniqueness, stability, and reconstruction of the conductivity from boundary measurements. A detailed distinction between the isotropic and anisotropic case is made, pointing out the major difficulties with the anisotropic case. The issues of global and local measurements are studied, noting that local measurements are more appropriate for practical applications such as screening for breast cancer.

1 Introduction

Electrical impedance tomography (EIT) is the recovery of the conductivity (or conductivity and permittivity) of the interior of a body from a knowledge of currents and voltages applied to its surface. In geophysics, where the method is used in prospecting and archaeology, it is known as electrical resistivity tomography. In industrial process tomography it is known as electrical resistance tomography or electrical capacitance tomography. In medical imaging, when at the time of writing it is still an experimental technique rather than routine clinical practice, it is called EIT. A very similar technique is used by weakly electric fish to navigate and locate prey and in this context it is called electrosensing. An example of a medical application of EIT is given in Fig. 1, which shows a 10-day-old healthy neonate breathing spontaneously and lying in the prone position with the head turned to the left. Sixteen EIT electrodes are placed in a transverse plane around the chest, and EIT data acquired with the Goe MF-II system. This child was a subject in a study which used EIT to examine patterns of breathing in neonates and the relationship to body posture [57]. In this study, EIT was able to show, for the first time, that, in a prone position, the lung on the opposite side (contralateral) to the face receives significantly larger air flows.

The simplest mathematical formulation of inverse problem of EIT can be stated as follows. Let Ω be a conducting body described by a bounded domain in $\mathbb{R}^n$, $n \geq 2$, with electrical conductivity a bounded and positive function $\gamma(x)$ (later complex γ will be considered). In the absence of internal sources, the electrostatic potential u in Ω is governed by the elliptic partial differential equation

$$L_\gamma u := \nabla \cdot \gamma \nabla u = 0 \qquad \text{in} \quad \Omega. \tag{1}$$

It is natural to consider the weak formulation of (1) in which $u \in H^1(\Omega)$ is a weak solution to (1). Given a potential $\phi \in H^{1/2}(\partial\Omega)$ on the boundary, the induced potential $u \in H^1(\Omega)$ solves the Dirichlet problem

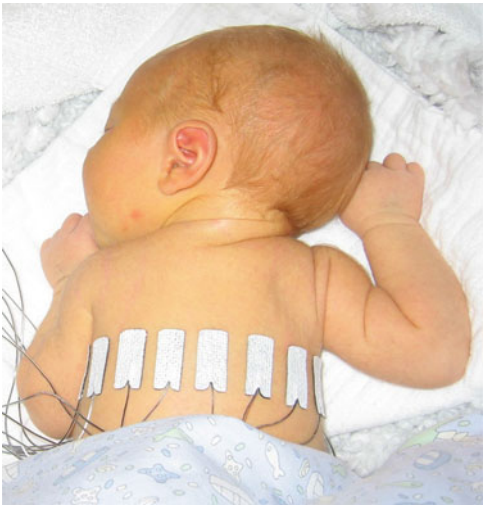


Fig. 1 10-day-old spontaneously breathing neonate lying in the prone position with the head turned to the left. Sixteen medical grade Ag/AgCl electrodes were placed in a transverse plane and connected to a Geo MF-II EIT system [57]

$$\begin{cases} L_\gamma u = 0 & \text{in} \quad \Omega, \\ u|_{\partial\Omega} = \phi. \end{cases} \tag{2}$$

The currents and voltages measurements taken on the surface of Ω, $\partial\Omega$, are given by the so-called Dirichlet-to-Neumann map (associated with γ) or voltage-to-current map

$$\Lambda_\gamma : u|_{\partial\Omega} \in H^{1/2}(\partial\Omega) \longrightarrow \gamma \frac{\partial u}{\partial \nu} \in H^{-1/2}(\partial\Omega).$$

Here, ν denotes the unit outer normal to $\partial\Omega$ and the restriction to the boundary is considered in the sense of the trace theorem on Sobolev spaces. Here, $\partial\Omega$ must be at least Lipschitz continuous and $\gamma \in L^\infty(\Omega)$ with $\operatorname{ess\,inf} \operatorname{Re} \gamma = m > 0$.

The *forward problem* under consideration is the map $\gamma \in \mathcal{D}_m \mapsto \Lambda_\gamma$, where $\mathcal{D}_m = \{\gamma \in L^\infty(\Omega) | \operatorname{ess\,inf} \gamma \geq m\}$ The *inverse problem* for complete data is then the recovery of γ from Λ_γ. As is usual in inverse problems, consideration will be given to questions of (1) uniqueness of solution (or from a practical point of view sufficiency of data), (2) stability/instability with respect to errors in the data, and (3) practical algorithms for reconstruction. It is also worth pointing out to the reader who is not very familiar with EIT the well-known fact that the behavior of materials under the influence of external electric fields is determined not only by the electrical conductivity γ but also by the electric permittivity ε so that the determination of the complex valued function $\gamma(x, \omega) = \sigma(x) + \mathrm{i}\omega\varepsilon(x)$ would be the more general and realistic problem, where $\mathrm{i} = \sqrt{-1}$ and ω is the frequency. The simple case where $\omega = 0$ will be treated in this work. For a description of the formulation of the inverse problem for the complex case, refer for example to [20]. Before addressing questions (1)–(3) mentioned above, it is interesting to consider how the problem arises in practice.

Measurement Systems and Physical Derivation

For the case of direct current, that is the voltage applied is independent of time, the derivation is simple. Of course here $\Omega \subset \mathbb{R}^3$. First suppose that it is possible to apply an arbitrary voltage $\phi \in H^{1/2}(\Omega)$ to the surface. It is typical to assume that the exterior $\mathbb{R}^3 \backslash \Omega$ is an electrical insulator. An electric potential (voltage) u results in the interior and the current $\mathbf{J}$ that flows satisfies the continuum Ohm's law $\mathbf{J} = -\gamma \nabla u$; the absence of current sources in the interior is expressed by the continuum version of Kirchoff's law $\nabla \cdot \mathbf{J} = 0$ which together result in (1). The boundary conditions are controlled or measured using a system of conducting electrodes which are typically applied to the surface of the object. In some applications, especially geophysical, these may be spikes that penetrate the object, but it is common to model these as points on the surface. Most EIT systems are designed to apply a known current on (possibly a subset) or electrodes and measure the voltage that results on electrodes (again possibly a subset, in some cases disjoint from those carrying a nonzero current). In other cases it is a predetermined voltage applied to electrodes and the current measured; there being practical reasons determined by electronics or safety for choosing one over the other. In medical EIT applying known currents and measuring voltages is typical. One reason for this is the desire to limit the maximum current for safety reasons. In practice the circuit that delivers a predetermined current can only do so while the voltage required to do that is within a certain range so both maximum current and voltage are limited. For an electrode (let us say indexed by ℓ) not modeled by a point but covering a region $E_\ell \subset \partial\Omega$ the current to that electrode is the integral

$$I_\ell = \int_{E_\ell} -\mathbf{J} \cdot \nu \, dx. \tag{3}$$

Away from electrodes,

$$\gamma \frac{\partial u}{\partial \nu} = 0, \qquad \text{on } \partial\Omega \backslash \bigcup_{\ell=1}^{L} E_\ell \tag{4}$$

as the air surrounding the object is an insulator. On the conducting electrode, $u|_{E_\ell} = V_\ell$ a constant, or as a differential condition

$$\nu \times \nabla u = 0 \qquad \text{on } \partial\Omega \backslash \bigcup_{\ell=1}^{L} E_\ell. \tag{5}$$

Taken together, (3)–(5) are called the *shunt model*. This ideal of a perfectly conducting electrode is of course only an approximation. Note that while the condition $u \in H^1(\Omega)$ is a sensible condition, which ensures finite total dissipated power, it is not sufficient to ensure (3) is well defined. Indeed for smooth γ and

smooth ∂E_ℓ the condition results in a square root singularity in the current density on the electrode. A more realistic model of electrodes is given later.

It is more common to use alternating current in geophysical and process monitoring applications, and essential in medical applications. Specifically the direction of the current must be reversed within a sufficiently short time to avoid electrochemical effects. This also means that the time average of the applied current should be zero. In medical applications, current in one direction for sufficient duration would result in transport of ions, and one of the effects of this can be stimulation of nerves. It would also degrade electrode behavior due to charge build up and ionic changes in the electrode. As a general rule higher levels of current and voltage are considered safer at higher temporal frequencies. The simplest EIT system therefore operates at a fixed frequency using an oscillator or digital signal processing to produce a sinusoidal current. Measurements are then taken of the magnitude, or in some cases the components that are in phase and $\pi/2$ out of phase with the original sine wave. When an EIT system switches from stimulating one set of electrodes to the next set in a stimululation pattern, the measurements adapt to the new pattern over a finite time, and typical EIT systems are designed to start measuring after this transient term has decayed so as to be negligible.

In geophysics a technique that is complementary to EIT called *induced polarization tomography* IPT is used to find polarizable minerals. In effect this uses a square wave pulse and measures the transient response [85]. In process tomography a technique known as electrical capacitance tomography is designed for imaging insulating materials with different dielectric permittivities, for example oil and gas in a pipe [60, 101]. Again square waves or pulses are used.

In medical and geophysical problems the response of the materials may vary with frequency. For example in a biological cell higher frequency current might penetrate a largely capacitive membrane and so be influenced by the internal structures of the cell, while lower frequency currents pass around the cell. This has led to electrical impedance tomography spectroscopy (EITS) [47], and in geophysics spectral induced polarization tomography (SIPT) [85]. The spectral response can be established either by using multiple sinusoidal frequencies or by sampling the transient response to a pulse.

Our starting point for the case of alternating current is the time harmonic Maxwell equations at a fixed angular frequency ω. Here it is assumed that the transient components of all fields are negligible and represent the time harmonic electric and magnetics vector fields using the complex representation $\mathcal{F}(x,t) = \mathrm{Re}\,(\mathbf{F}\exp(\mathrm{i}\omega t))$, yielding

$$\nabla \times \mathbf{E} = -\mathrm{i}\omega\mathbf{B}, \tag{6}$$

$$\nabla \times \mathbf{H} = \mathbf{J} + \mathrm{i}\omega\mathbf{D}. \tag{7}$$

The electric and magnetic fields $\mathbf{E}$ and $\mathbf{H}$ are related to the current density $\mathbf{J}$, electric displacement $\mathbf{D}$, and magnetic flux $\mathbf{B}$ by the material properties conductivity σ, permittivity ϵ, and permeability μ by

$$\mathbf{J} = \sigma\mathbf{E},\ \mathbf{D} = \epsilon\mathbf{E},\ \mathbf{B} = \mu\mathbf{H}. \tag{8}$$

The fields **E** and **H** are evaluated on directed curves, while the "fluxes" **J**, **D** and **B** on surfaces. In biomedical applications one can typically take μ to be constant and to be the same inside the body as outside in air. In non-destructive testing and geophysical applications there may well be materials with differing permeability. Here (8) assumes linear relations. For example the first is the continuum Ohm's law. Here, the material properties may be frequency dependent. This *dispersion* is important in EIS and SIPT. For the moment a first approximation is to assume isotropy (so that the material properties are scalars).

There are many inverse problems governed by time harmonic Maxwell's equations.. These occur at large values of ω and include optical and microwave tomographic techniques and scattering problems such as radar which are not discussed in this chapter. There are also systems where the fields arise from alternating current in a coil, and measurements are made either with electrodes or with other coils. Mutual (or magnetic) induction tomography (MIT) falls into this category and has been tried in medical and process monitoring applications [46]. In these cases the eddy current approximation [9] to Maxwell's equations is used. While for direct current EIT (i.e., ERT) the object is assumed surrounded by an insulator, in MIT one must account for the magnetic fields in the surrounding space, there being no magnetic "shielding."

The assumptions used to justify the usual mathematical model of EIT make it distinct from many other inverse problems for Maxwell's equations. Given

Assumption 1. Transients components of all fields are negligible.

This assumption simply means that a sufficient "settling time" has been given before making measurements.

The interest in relatively low frequencies, where magnetic effects can be neglected, translates into two assumptions

Assumption 2. $\omega\sqrt{\epsilon\mu}$ is small compared with the size of Ω.

This means that the wavelength of propagating waves in the material is large. A measurement accuracy of $2^{-12} = 1/4096$ is ambitious at higher frequencies means that for wave effects to be negligible

$$d\,\omega\sqrt{\epsilon\mu} < \cos^{-1}\frac{4095}{4096}, \tag{9}$$

where d is the diameter of the body. Taking the relative permittivity to be 10 and $R = 0.3$ m gives a maximum frequency of 1 MHz.

Assumption 3. $\sqrt{\omega\sigma\mu/2}$ is small compared with the size of Ω.

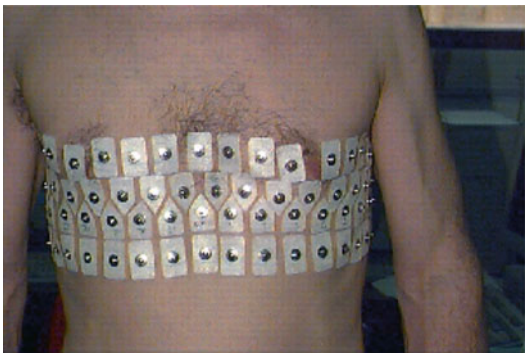


Fig. 2 A system of electrodes used for chest EIT at Oxford Brookes University. The positions of the electrodes were measured manually with a tape measure and the cross-sectional shape was also determined by manual measurements. These electrodes have a disk of jell containing silver chloride solution that makes contact with the skin. Each electrode was attached to the EIT system by a screened lead, not shown in this picture for clarity

The quantity

$$\delta = \sqrt{\frac{2}{\omega\sigma\mu}} \tag{10}$$

is known as the skin depth. For a frequency of 10 kHz and a conductivity of $0.5\,\mathrm{Sm}^{-1}$ typical in medical applications, the skin depth is 7 m. In geophysics lower frequencies are typical but length scales are larger. In a conducting cylinder the electric field decays with distance r from the boundary at a rate $\mathrm{e}^{-r/\delta}$ due to the opposing magnetic field. At EIT frequencies this simple example suggests that accurate forward modeling of EIT should take account of this effect although it is currently not thought to be a dominant source of error.

The effect of Assumptions 2 and 3 combined together is that it is reasonable to neglect $\nabla \times E$ in Maxwell's equations resulting in the standard equation for complex EIT

$$\nabla \cdot (\sigma + \mathrm{i}\omega\epsilon)\nabla u = 0. \tag{11}$$

Here the expression $\gamma = \sigma + \mathrm{i}\omega\epsilon$ is called complex conductivity, or logically the *admittivity*, while $1/\sigma$ is called *resistivity* and the rarely-used complex $1/\gamma$ *impedivity*. A scaling argument is given for the approximation (11) in [31], and numerical checks on the validity of the approximation in [37, 104] (Fig. 2).

It is often not so explicitly stated but while in the direct current case one can neglect the conductivity of the air surrounding the body, for the alternating current case the electrodes are coupled capacitively and, while σ can be assumed to be

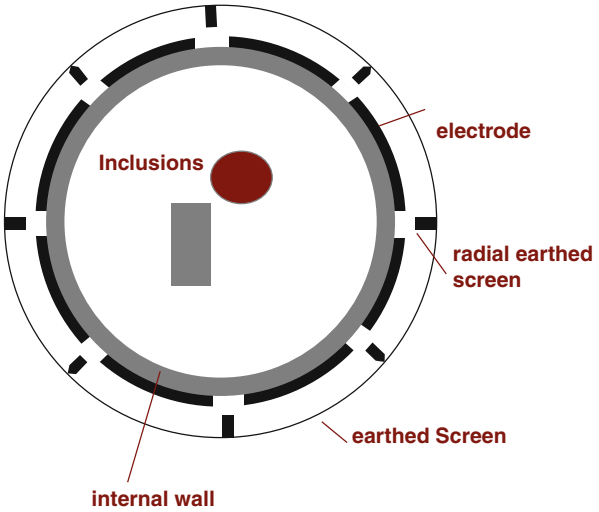


Fig. 3 A cross section through a typical ECT sensor around a pipe (internal wall) showing the external screen with radial screens designed to reduce the external capacitive coupling between electrodes

zero for air, the permittivity of any material is no smaller than that of a vacuum $\epsilon_0 = 8.85 \times 10^{-12}$, although dry air approaches that value. One requires then

Assumption 4. $\omega\epsilon$ in the exterior is negligible compared to $|\sigma + i\omega\epsilon|$ in the interior.

For example, with a conductivity or $0.2\,\mathrm{Sm}^{-1}$, the magnitude of the exterior admittivity reaches 2^{-12} of that value for a frequency of 0.88 MHz. For a more detailed calculation the capacitance between the electrodes externally could be compared with the impedance between electrodes. In ECT frequencies above 1 MHz are used and the exterior capacitance cannot be neglected. Indeed an exterior grounded shield is used so that the exterior capacitive coupling is not affected by the surroundings (see Fig. 3).

The Concentric Anomaly: A Simple Example

A simple example helps us to understand the instability in the inverse conductivity problem. Let Ω be the unit disk in $\mathbb{R}^2$ with polar coordinates (r, θ) and consider a concentric anomaly in the conductivity of radius $\rho < 1$

$$\gamma(x) = \begin{cases} a_1, & |x| \le \rho \\ a_0, & \rho < |x| \le 1. \end{cases} \tag{12}$$

From separation of variables, matching Dirichlet and Neumann boundary conditions at $|x| = \rho$, for $n \in \mathbb{Z}$

$$\Lambda_\gamma e^{in\theta} = |n| \frac{1 + \mu\rho^{2|n|}}{1 - \mu\rho^{2|n|}} e^{in\theta}, \tag{13}$$

where $\mu = (a_1 - a_0)/(a_1 + a_0)$. From this one sees the effect of the Dirichlet to Neumann map on the complex Fourier series, and the effect on the real Fourier series is easily deduced. This example was considered in [64] as an example of the eigenvalues and eigenfunctions of Λ_γ, and also by [2] as an example of instability. Thus $||\gamma - a_0||_{L^\infty(\Omega)} = |a_1 - a_0|$ independently of ρ and yet $\Lambda_\gamma \to \Lambda_{a_0}$ in the operator norm. Hence if an inverse map $\Lambda_\gamma \mapsto \gamma$ exists, it cannot be continuous in this topology. Similar arguments can be used to show instability of inversion in other norms.

This example reveals many other features of the more general problem. For example experimentally one observes *saturation*: for an object placed away from the boundary, changes in the conductivity of an object with a conductivity close to the background are fairly easily detected, but for an object of very high or low conductivity further changes in conductivity of that object have little effect. This saturation effect was explored for offset circular objects (using conformal mappings) by Seagar [98]. This is also an illustration of the nonlinearity of $\gamma \to \Lambda_\gamma$. One can also see in this example that smaller objects (with the same conductivity) produce smaller changes in measured data as one might expect.

On the unit circle S^1 one can define an equivalent norm on the Sobolev space $H^s_\diamond(S^1)$ (see definitions in the section "The Neumann-to-Dirichlet Map") by

$$\left\| \sum_{n=-\infty, n\neq 0}^{\infty} c_n m r e^{in\theta} \right\|_s^2 = \sum_{n=-\infty, n\neq 0}^{\infty} n^{2s} c_n^2. \tag{14}$$

It is clear for this example that $\Lambda_\gamma : H^s_\diamond(S^1) \to H^{s-1}_\cdot(S^1)$, for any s. Roughly the current is a derivative of potential and one degree of differentiability less smooth. Technically Λ_γ (for any positive $\gamma \in C^\infty(\Omega)$) is a *first order pseudo-differential operator* [80]. The observation that for the example $e^{-in\theta}\Lambda_\gamma e^{in\theta} = |n| + o(n^{-p})$ as $|n| \to \infty$ for any $p > -1$ illustrates that the change in conductivity and radius of the interior object is of somewhat secondary importance! In the language of pseudodifferential operators for a general γ such that $\gamma - 1$ vanishes in a neighborhood of the boundary, Λ_γ and Λ_1 *differ by a smoothing operator*.

Since (13), Λ_γ^{-1} is also a well-defined operator on $L^2_\cdot \to L^2_\cdot$ with eigenvalues $O(|n|^{-1})$ and is therefore a Hilbert–Schmidt operator. This is also known for the general case [36].

Early work on medical applications of EIT [58, 74] hoped that the forward problem in EIT would be approximated by generalized ray transform – that is integrals along current stream lines. The example of a concentric anomaly was used to illustrate that EIT is *nonlocal* [99]. If one applies the voltage $\cos(\theta + \alpha)$, which for a homogeneous disk would result in current streamlines that are straight and parallel, a change in conductivity in a small radius ρ from the center changes

all measured currents, not just on lines passing through the region of changed conductivity $|x| \leq \rho$. In the 1980s a two-dimensional algorithm that backprojected filtered data along equipotential lines was popularized by Barber and Brown [11]. Berenstein [17] later showed that the linearized EIT problem in a unit disc can be interpreted as the Radon transform with respect to the Poincaré metric and a convolution operator and that Barber and Brown's algorithm is an approximate inverse to this.

In process applications of EIT and related techniques the term *soft field imaging* is used, which by analogy to soft field X-rays means a problem that is nonlinear and nonlocal. However in the literature when the "soft field effect" is invoked, it is often not clear if it is the nonlinear or nonlocal aspect to which they refer and, in the authors' opinion, the term is best avoided.

Measurements with Electrodes

A typical electrical imaging system uses a system of conducting electrodes attached to the surface of the body under investigation. One can apply current or voltage to these electrodes and measure voltage or current respectively. For one particular measurement the voltages (with respect to some arbitrary reference) are V_ℓ and the currents I_ℓ, which are arranged in vectors as $\mathbf{V}$ and $\mathbf{I} \in \mathbb{C}^L$. The discrete equivalent of the Dirichlet-to-Neumann Λ_γ map is the transfer admittance, or mutual admittance matrix $\mathbf{Y}$ which is defined by $\mathbf{I} = \mathbf{YV}$.

It is easy to see that the vector $\mathbf{1} = (1, 1, \ldots, 1)^T$ is in the null space of $\mathbf{Y}$, and that the range of $\mathbf{Y}$ is orthogonal to the same vector. Let S be the subspace of $\mathbb{C}^L$ perpendicular to $\mathbf{1}$; then it can be shown that $\mathbf{Y}|_S$ is invertible from S to S. The generalized inverse (see chapter ▸ Linear Inverse Problems) $\mathbf{Z} = \mathbf{Y}^\dagger$ is called the transfer impedance. This follows from uniqueness of solution of *shunt model* boundary value problem.

The transfer admittance, or equivalently transfer impedance, represents a complete set of data which can be collected from the L electrodes at a single frequency for a stationary linear medium. It can be seen from the weak formulation of (11) that $\mathbf{Y}$ and $\mathbf{Z}$ are symmetric (but for $\omega \neq 0$ *not Hermittian*). In electrical engineering this observation is called *reciprocity*. The dimension of the space of possible transfer admittance matrices is clearly no bigger than $L(L-1)/2$, and so it is unrealistic to expect to recover more unknown parameters than this. In the analogous case of planar resistor networks with L "boundary" electrodes the possible transfer admittance matrices can be characterized completely [33], a characterization which is known at least partly to hold in the planar continuum case [63]. A typical electrical imaging system applies current or voltage patterns which form a basis of the space S, and measures some subset of the resulting voltages which as they are only defined up to an additive constant can be taken to be in S.

The shunt model is nonphysical; in medical application with electrodes applied to skin and in "phantom" tanks used to test EIT systems with ionic solutions in contact with metal electrodes, a contact impedance layer exists between the solution or skin

and the electrode. This modifies the shunting effect so that the voltage under the electrode is no longer constant. The voltage on the electrode is still a constant V_ℓ so now on E_ℓ there is a voltage drop across the contact impedance layer

$$\phi + z_\ell \sigma \frac{\partial \phi}{\partial \nu} = V_\ell, \tag{15}$$

where the contact impedance z_ℓ could vary over E_ℓ but is usually assumed constant. Experimental studies have shown [56] that a contact impedance on each electrode is required for an accurate forward model. This new boundary condition together with (3) and (4) form the *Complete Electrode Model* or CEM. For experimental validation of this model see [30], theory [103], and numerical calculations [96,111]. A nonzero contact impedance removes the singularity in the current density, although high current densities still occur at the edges of the electrodes (Fig. 4). For further details on the singularity in the current density, see [32]. While "point electrodes," in which the current density is a sum of delta functions, are a limiting case of the CEM, they are not physically realistic as they result in nonphysical potentials not in $H^1(\Omega)$. The trace on the boundary cannot be evaluated at a point, so point measurements of voltage are undefined. However it can be shown that if the conductivity is changed only in the compliment of a neighborhood of $\partial\Omega$ the resulting *voltage difference* at the boundary can be evaluated at points [52].

The set of imposed current patterns, or *excitation patterns*, is designed to span S, or at least that part of it that can be accurately measured in a given situation. In medical EIT, with process ERT following suit, early systems designed at Sheffield [11] assumed a two-dimensional circular domain. Identical electrodes were equally spaced on the circumference and, taking them to be numbered anticlockwise, the excitation patterns used were adjacent pairs, that is proportional to

$$I_\ell^i = \begin{cases} 1, & i = \ell \\ -1, & i = \ell + 1 \\ 0, & \text{otherwise}, \end{cases} \tag{16}$$

for $i = 1, \ldots L-1$. The electronics behind this is balanced current source connected between two electrodes [59, Chap. 2], and this is somewhat easier to achieve in practice than a variable current source at more than two electrodes. For general geometries, where the electrodes are not placed on a closed curve, other pairs of electrodes are chosen. For example $I_1^i = -1$, while $I_\ell^i = \delta_{i\ell}, \ell \neq 1$.

Measurements of voltage can only be differential and so voltage measurements are taken between pairs of electrodes, for example adjacent pairs, or between each and some fixed electrode. In pair drive systems, similar to the original Sheffield system, voltages on electrodes with nonzero currents are not measured, resulting in incomplete knowledge of $\mathbf{Z}$.

In geophysical surface resistivity surveys it is common to use a pair drive and pair measurement system, using electrodes in a line where a two-dimensional approximation is used, or laid out in a rectangular or triangular grid where the

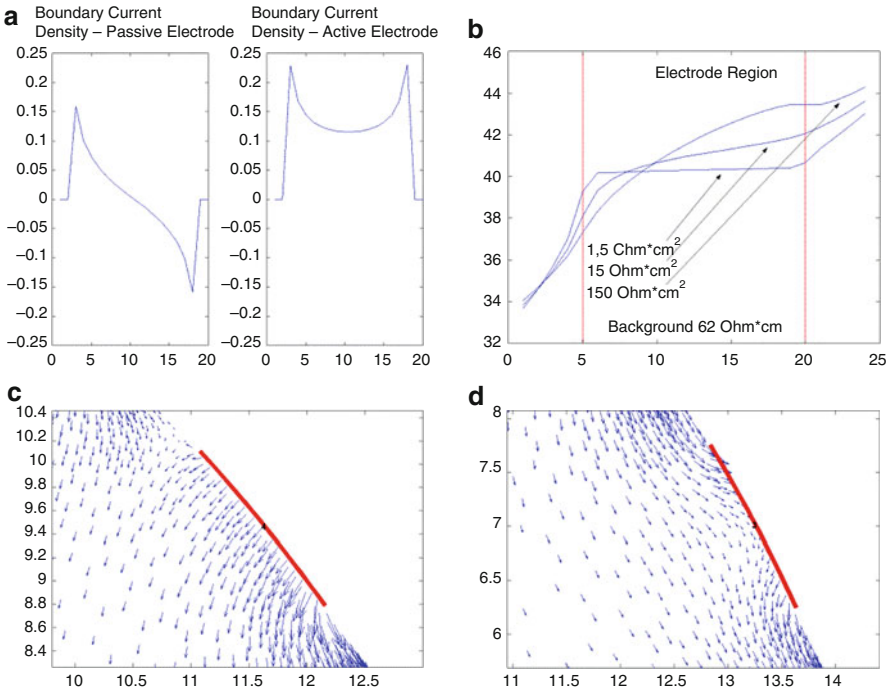


Fig. 4 The current density on the boundary with the CEM is greatest at the edge of the electrodes, even for passive electrodes. This effect is reduced as the contact impedance increases. Diagrams courtesy of Andrea Borsic. (**a**) Current density on the boundary for passive and active electrodes. In fact there is a jump discontinuity at the edge of electrodes for nonzeros contact impedance although the plotting routine has joined the *left* and *right* limits. (**b**) The effect of contact impedance on the potential beneath an electrode. The potential *is* continuous. (**c**) Interior current flux near an active electrode. (**d**) Interior current flux near a passive electrode

full three-dimensional problem is solved. Measurements taken between pairs of non-current carrying electrodes. The choice of measurement strategy is limited by the physical work involved in laying out the cables and by the switching systems. Often electrodes will be distributed along one line and a two-dimensional approximate reconstruction is used as this gives adequate information for less cost. A wider spacing of the current electrodes is used where the features of interest is located at a greater depth below the ground. In another geophysical configuration, cross borehole tomography, electrodes are deployed down several vertical cylindrical holes in the ground, typically filled with water, and current passed between electrodes in the same or between different bore holes. Surface electrodes may be used together with those in the bore holes. In some systems the current is measured to account for a non-ideal current source.

In capacitance tomography a basis of voltage patterns is applied and the choice $V_\ell^i = \delta_{i\ell}$ is almost universal. The projection of these vectors to S (denoted as an "electrode-wise basis") is convenient computationally as a current pattern.

Given a *multiple drive system* capable of driving an arbitrary vector of currents in S (in practice with in some limits on the maximum absolute current and on the maximum voltage) there remains a choice of excitation patterns. While exact measurements of $\mathbf{ZI}^i$ for $\mathbf{I}^i$ in any basis for S is clearly sufficient, the situation is more complicated with measurements of finite precision in the presence of noise. If a redundant set of currents is taken, the problem of estimating $\mathbf{Z}$ becomes one of *multivariate linear regression*. The choice of current patterns is then a *design matrix*. Another approach seeks the minimum set of current patterns that results in usable measurements. Applying each current pattern and taking a set of measurements take a finite time, during which the admittivity changes. Without more sophisticated statistical methods (such as Kalman filters [112]), there are diminishing returns in applying redundant current patterns. Suppose that the total power $\mathbf{V}^*\mathbf{ZI}$ is constrained (to keep the patient electrically safe) and the current best estimate of the admittivity gives a transfer admittance $\mathbf{Z}_{\text{calc}}$, then it is reasonable to apply currents $\mathbf{I}$ such that $(\mathbf{Z} - \mathbf{Z}_{\text{calc}})\mathbf{I}$ is above the threshold of voltages that can be accurately measured and modeled. One approach is to choose current patterns that are the right generalized singular vectors of $\mathbf{Z} - \mathbf{Z}_{\text{calc}}$ with singular values bigger than an error threshold. The generalized singular values are with respect to the norm $||\mathbf{I}||_{\mathbf{Z}} := ||\mathbf{ZI}||$ on S and are the extrema of the *distinguishability* defined as

$$\frac{||(\mathbf{Z} - \mathbf{Z}_{\text{calc}})\mathbf{I}||}{||\mathbf{I}||_{\mathbf{Z}}}, \tag{17}$$

for $\mathbf{I} \in S$. These excitation patterns are called "optimal current patterns" [45] and can be calculated from an iterative procedure involving repeated measurement. For circular disk with rotationally symmetric admittivity and equally spaced identical electrodes, the singular vectors will be discrete samples of a Fourier basis and these *trigonometric patterns* are a common choice for multiple drive systems using a circular array of electrodes.

2 Uniqueness and Stability of the Solution

Uniqueness of solution is very important in inverse problems, although when talking to engineers it is often better to speak of *sufficiency of data* to avoid confusion. Interestingly it is generally true that results that show *insufficiency of data*, that one cannot recover an unknown function even if an infinite number of measurements of arbitrary precision are taken, have more impact in applied areas. While there are still unsolved problems in the uniqueness theory for the EIT inverse problem, there has been considerable progress over the last three decades and many important questions have been answered. While for an isotropic real conductivity γ (with certain smoothness assumptions for dimensions $n \geq 3$), γ is uniquely determined by the complete data Λ_γ (see [10, 24, 107]), an anisotropic conductivity tensor is not uniquely determined by the boundary data, although some progress on what can be determined in this case has been made (see [3, 6, 43, 79]). Aside from

knowing what can and cannot be determined with ideal data, there are two important ways the theoretical work has a practical impact. Firstly in some cases the proof of uniqueness of solution suggests a reconstruction algorithm. As for the two-dimensional case (below) the most effective approach (the so-called $\bar{\partial}$-method) to uniqueness theory has now been implemented as a fast, practical algorithm. The other is an understanding of the instability and conditional stability of the inverse problem. This helps us to determine what a priori information is helpful in reducing the sensitivity of the solution to errors in the data.

In 1980 Calderón published a paper with the title "On a inverse boundary value problem" [26], where he addressed the problem of whether it is possible to determine the conductivity of a body by making current and voltage measurements at the boundary. It seems that Calderón thought of this problem when he was working as an engineer in Argentina for the Yacimientos Petroliféros Fiscales (YPF), but it was only decades later that he decided to publish his results. This short paper is considered the first mathematical formulation of the problem. For a reprinted version of this manuscript, refer to [27]. The authors wish to recall also the work due to Druskin (see [38–40]) which has been carried on independently from Calderón's approach and has been devoted to the study of the problem from a geophysical point of view.

The Isotropic Case

Calderón's Paper

Calderón considered a domain Ω in $\mathbb{R}^n$, $n \geq 2$, with Lipschitz boundary $\partial\Omega$. He took γ be a real bounded measurable function in Ω with a positive lower bound. Let Q_γ be the quadratic form (associated with Λ_γ) defined by

$$Q_\gamma(\phi) = \langle \phi, \Lambda_\gamma \phi \rangle = \int_\Omega \gamma \, |\nabla u|^2 \, dx, \tag{18}$$

where $u \in H^1(\Omega)$ solves the Dirichlet problem (2). Physically $Q_\gamma(\phi)$ is the Ohmic power dissipated when the boundary voltage ϕ is applied. The bilinear form associated with Q_γ is then obtained by using the polarization identity

$$\begin{aligned} B_\gamma(\phi, \psi) &= \frac{1}{2}\Big\{ Q_\gamma(\phi + \psi) - Q_\gamma(\phi) - Q_\gamma(\psi) \Big\} \\ &= \frac{1}{2}\Big\{ \int_\Omega \left(\gamma |\nabla(u+v)|^2 - \gamma |\nabla u|^2 - \gamma |\nabla v|^2 \right) dx \Big\} \\ &= \int_\Omega \gamma \nabla u \cdot \nabla v \, dx, \end{aligned} \tag{19}$$

where $L_\gamma v = 0$ in Ω and $v|_{\partial\Omega} = \psi \in H^{\frac{1}{2}}(\partial\Omega)$. Clearly a complete knowledge of any of Λ_γ, Q_γ and B_γ are equivalent. Calderón considered the "forward" map

$$\mathbf{Q}: \gamma \longrightarrow Q_\gamma$$

and proved that $\mathbf{Q}$ is bounded and analytic in the subset of $L^\infty(\Omega)$ consisting of functions γ which are real and have a positive lower bound. He then investigated the injectivity of the map and in order to do so, he linearized the problem. He in fact proved the injectivity of the Fréchet derivative of $\mathbf{Q}$ at $\gamma = 1$. Here, a few details of the linearization for a general γ are given. Let u be the solution to (2) and $U = u + w$ satisfy $L_{\gamma+\delta}U = 0$, with $U|_{\partial\Omega} = \phi$. The perturbation in potential satisfies $w|_{\partial\Omega} = 0$, by considering the Dirichlet data fixed and investigating how the Neumann data varies when γ is perturbed to $\gamma + \delta$. This yields

$$L_\delta u + L_\gamma w + L_\delta w = 0. \tag{20}$$

Now let $G : H^{-1}(\Omega) \to H_0^1(\Omega)$ be the Green's operator that solves the equivalent of Poisson's equation for L_γ with zero Dirichlet boundary conditions. That is for $g \in H^{-1}(\Omega)$, $L_\gamma G g = g$ and $G(g)|_{\partial\Omega} = 0$, the operator equation is given

$$(1 + GL_\delta)w = -GL_\delta u. \tag{21}$$

An advantage of using the L^∞ norm is that it is clear $||L_\delta|| \to 0$ in the $H^1(\Omega) \to H^{-1}(\Omega)$ operator norm as $||\delta||_\infty \to 0$. This means one can choose δ small enough that $||GL_\delta|| < 1$ (in the operator norm on $H^1(\Omega)$) and this ensures that the term in the bracket in (21) is invertible and the operator series in

$$w = -\left(\sum_{k=1}^{\infty}(-GL_\delta)^k\right)u \tag{22}$$

is convergent. This proves that the map $\gamma \mapsto u$ and hence $\mathbf{Q}$ is not just C^∞ but analytic with (22) its Taylor series. Thus, the linearization of the map $\gamma \mapsto \Lambda_\gamma$ is

$$\Lambda_{\gamma+\delta}\phi = \Lambda_\gamma\phi + \gamma\frac{\partial}{\partial\nu}GL_\delta u + \delta\frac{\partial u}{\partial\nu} + O(||\delta||_\infty^2). \tag{23}$$

A strength of this argument is that it gives the Fréchet derivative in these norms, rather than just the Gateaux derivative. It is easy to deduce that the Fréchet derivative of $\mathbf{Q}$ at γ in the direction δ is given by

$$d\mathbf{Q}(\gamma)\delta(\phi) = \int_\Omega \delta|\nabla u|^2\, dx. \tag{24}$$

In many practical situations it is more common to fix the Neumann boundary conditions and measure the change in boundary voltage as the conductivity changes. Suppose $L_\gamma u = 0$, $L_{\gamma+\delta} U = 0$, $w = U - u$ with

$$\gamma \partial u/\partial \nu = (\gamma + \delta)\partial U/\partial \nu = g \in H_\cdot^{-1/2}(\partial\Omega)$$

then a similar argument to the above shows

$$\int_{\partial\Omega} w\gamma \frac{\partial u}{\partial \nu}\, dx = -\int_\Omega \delta |\nabla u|^2\, dx + O(||\delta||_\infty^2). \tag{25}$$

The polarization identity is often applied to (25) giving

$$\int_{\partial\Omega} w\gamma \frac{\partial v}{\partial \nu}\, dx = -\int_\Omega \delta \nabla u \cdot \nabla v\, dx + O(||\delta||_\infty^2), \tag{26}$$

where $L_\gamma v = 0$. This is often used in practice with

$$\gamma \frac{\partial v}{\partial \nu} = \chi_{E_i}/|E_i| - \chi_{E_j}/|E_j|, \tag{27}$$

which represents the difference in the characteristic functions of a pair of electrodes. In the case of the shunt model this makes the left-hand side of (25) equal to change that occurs in the difference between voltages on that pair of electrodes when the conductivity is perturbed. The formula (25) and its relatives are referred to as the Geselowitz Sensitivity Theorem in the bioengineering literature. With the CEM (25) still holds, but with u and v satisfying (15) [97].

Returning to Calderón's argument: for $\gamma = 1$ one has $L_1 u = \nabla^2 u$. To prove the injectivity of $d\mathbf{Q}(1)$ one must show that if the integral appearing in (24) vanishes for all the harmonic functions in Ω, then $\delta = 0$ in Ω. Suppose the integral in (24) vanishes for all $u \in H^1(\Omega)$ such that $\nabla^2 u = 0$ in Ω, then

$$\int_\Omega \delta \nabla u \cdot \nabla v = 0, \tag{28}$$

whenever $\nabla^2 u = \nabla^2 v = 0$ in Ω. For any $z \in \mathbb{R}^n$ consider $a \in \mathbb{R}^n$ such that $|a| = |z|$, $a \cdot z = 0$ and consider the harmonic functions

$$\begin{aligned} u(x) &= \mathrm{e}^{\pi i (z\cdot x) + \pi (a \cdot x)}, \\ v(x) &= \mathrm{e}^{\pi i (z\cdot x) - \pi (a \cdot x)}, \end{aligned} \tag{29}$$

which is equivalent to choosing

$$u(x) = \mathrm{e}^{x \cdot \rho}, \qquad v(x) = \mathrm{e}^{-x \cdot \bar{\rho}},$$

where $\rho \in \mathbb{C}^n$ with

$$\rho \cdot \rho = 0.$$

Here the real dot product on complex vectors is used $\rho \cdot \rho := \rho^T \rho$. With the choice made in (29), (28) leads to

$$2\pi|z|^2 \int \delta(x) \mathrm{e}^{2\pi i(z^\star \cdot x)}\, dx = 0, \qquad \text{for each} \quad z,$$

therefore $\delta(x) = 0$, for all $x \in \Omega$. Calderón also observed that if the linear operator $d\mathbf{Q}(1)$ had a closed range, then one could have concluded that Q itself was injective in a sufficiently small neighborhood of γ=constant. However conditions on the range of $d\mathbf{Q}(1)$, that would allow us to use the implicit function theorem, are either false or not known. Furthermore if the range was closed, one could have also concluded that the inverse of $d\mathbf{Q}(1)$ was a bounded linear operator by the open mapping theorem. Calderón concluded the paper by giving an approximation for the conductivity γ if

$$\gamma = 1 + \delta$$

and δ is small enough in the L^∞ norm, by making use of the same harmonic functions (29). Calderón's technique is based on the construction of low frequency oscillating solutions. Sylvester and Uhlmann proved in their fundamental paper [107] a result of uniqueness using high frequencies oscillating solutions of $L_\gamma u = 0$. Their solutions are of type

$$u(x, \xi, t) = \mathrm{e}^{x \cdot \xi}\, \gamma^{-\frac{1}{2}}(1 + \psi(x, \xi, t)),$$

which behaves (for high frequencies ξ) in the same way as the solutions used by Calderón. These oscillating solutions have come to be known as *complex geometrical optics (CGO) solutions*. Before going into more details of the use of CGO solutions, an earlier result using a different approach is given.

Uniqueness at the Boundary

In 1984 Kohn and Vogelius [75] proved that boundary values, and derivatives at the boundary, of a smooth isotropic conductivity γ could be determined from the knowledge of Q_γ. Their result is given by the following theorem.

Theorem 1. *Let Ω be a domain in $\mathbb{R}^n$ ($n \geq 2$) with smooth boundary $\partial\Omega$. Suppose $\gamma_i \in C^\infty(\bar{\Omega})$, $i = 1, 2$ is strictly positive and that there is a neighborhood B of some $x^\star \in \partial\Omega$ so that*

$$Q_{\gamma_1}(f) = Q_{\gamma_2}(f), \quad \textit{for all } f, \quad f \in H^{\frac{1}{2}}(\partial\Omega), \quad \mathrm{supp}(f) \subset B.$$

Then

$$\frac{\partial^{|\alpha|}}{\partial x^\alpha}\,\gamma_1(x^\star) = \frac{\partial^{|\alpha|}}{\partial x^\alpha}\,\gamma_2(x^\star), \qquad \forall \alpha.$$

Theorem 1 is a local result in the sense that one only need to know Q_γ in a open set of the boundary in order to determine the Taylor series of γ on that open set. The global reformulation of this result given in terms of Λ_γ is given below.

Theorem 2. *Let $\gamma_i \in C^\infty(\bar{\Omega})$, $i = 1, 2$ be strictly positive. If $\Lambda_1 = \Lambda_2$, then*

$$\frac{\partial^{|\alpha|}}{\partial x^\alpha}\,\gamma_1 = \frac{\partial^{|\alpha|}}{\partial x^\alpha}\,\gamma_2, \text{ on } \partial\Omega, \qquad \forall \alpha.$$

For a sketch of the proof of Theorem 2 see [110, Sketch of proof of Theorem 4.1, pp 6]. This result settled the identifiability question in the real-analytic category of conductivities. Kohn and Vogelius have extended this result to piecewise real-analytic (e.g., piecewise constant) conductivities in [76]. The proof of this result is based on [75] together with the Runge approximation theorem for solutions of $L_\gamma u = 0$.

CGO Solutions for the Schrödinger Equation

In 1987 Sylvester and Uhlmann [106, 107] constructed in dimension $n \geq 2$ CGO solutions in the whole space for the Schrödinger equation with potential q. Before giving their result, the well-known relation between the conductivity equation and the Schrödinger equation will be derived. This relationship is also important in diffuse optical tomography (see chapter ▸ Optical Imaging).

Lemma 1. *Let $\gamma \in C^2(\bar{\Omega})$ be strictly positive, yielding*

$$\gamma^{-\frac{1}{2}} L_\gamma(\gamma^{-\frac{1}{2}}) = \nabla^2 - q, \tag{30}$$

where

$$q = \frac{\nabla^2(\gamma^{\frac{1}{2}})}{\gamma^{\frac{1}{2}}}.$$

Proof of Lemma 1.

$$L_\gamma u = \gamma \nabla^2 u + \nabla\gamma \cdot \nabla u \tag{31}$$

therefore

$$\gamma^{-\frac{1}{2}} L_\gamma u = \gamma^{\frac{1}{2}} \nabla^2 u + \frac{\nabla\,\gamma \cdot \nabla u}{\gamma^{\frac{1}{2}}}.$$

Consider for $w = \gamma^{\frac{1}{2}} u$

$$\begin{aligned}
\nabla^2 w - q\, w &= \nabla^2 \left(\gamma^{\frac{1}{2}}\, u\right) - \left(\nabla^2 \gamma^{\frac{1}{2}}\right) u \\
&= \nabla \cdot \left(\nabla(\gamma^{\frac{1}{2}} u)\right) - \left(\nabla^2 \gamma^{\frac{1}{2}}\right) u \\
&= \nabla \cdot \left((\nabla \gamma^{\frac{1}{2}} u) + \gamma^{\frac{1}{2}} (\nabla u)\right) - \left(\nabla^2 \gamma^{\frac{1}{2}}\right) u \\
&= (\nabla^2 \gamma^{\frac{1}{2}}) u + 2 \nabla \gamma^{\frac{1}{2}} \cdot \nabla u + \gamma^{\frac{1}{2}} \nabla^2 u - (\nabla^2 \gamma^{\frac{1}{2}}) u \\
&= \gamma^{\frac{1}{2}} \nabla^2 u + \frac{\nabla \gamma \cdot \nabla u}{\gamma^{\frac{1}{2}}} \\
&= \gamma^{-\frac{1}{2}} L_\gamma u,
\end{aligned}$$

which proves (30). □

The term q is usually called the *potential* of the Schrödinger equation, by analogy with the potential energy in quantum mechanics, this definition being somehow confusing given that in EIT u is the electric potential. The results in [106, 107] state the existence of CGO solutions for the Schrödinger equation with potential q bounded and compactly supported in $\mathbb{R}^n$. This result is as given in [110], which relies on the weighted L^2 space $L^2_\delta(\mathbb{R}^n) = \{f : \int_{\mathbb{R}^n} (1+|x|^2)^\delta |f(x)|^2\, dx\}$. For $\delta < 0$ this norm controls the "growth at infinity." The Sobolev spaces $H^k_\delta(\mathbb{R}^n)$ are formed in the standard way from $L^2_\delta(\mathbb{R}^n)$

$$H^k_\delta(\mathbb{R}^n) = \{f \in W^k(\mathbb{R}^n) \mid D^\alpha f \in L^2_\delta(\mathbb{R}^n), \quad \text{for all } |\alpha| \le k\},$$

where α is a multi-index, $D^\alpha f$ denotes the αth weak derivative of f and $W^k(\mathbb{R}^n)$ is the set of k times weakly differentiable functions on $\mathbb{R}^n$.

Theorem 3. *Let $q \in L^\infty(\mathbb{R}^n)$, $n \ge 2$, with $q(x) = 0$ for $|x| \ge R > 0$ and $-1 < \delta < 0$. Then there exists $\epsilon(\delta)$ and such that for every $\rho \in \mathbb{C}^n$ satisfying*

$$\rho \cdot \rho = 0$$

and

$$\frac{||(1+|x|^2)^{1/2} q||_{L^\infty(\mathbb{R}^n)} + 1}{|\rho|} \le \epsilon$$

there exists a unique solution to

$$(\nabla^2 - q) u = 0 \tag{32}$$

of the form

$$u(x,\rho) = \mathrm{e}^{x\cdot\rho}\left(1+\psi_q(x,\rho)\right), \tag{33}$$

with $\psi_q(\cdot,\rho) \in L^2_\delta(\mathbb{R}^n)$. *Moreover* $\psi_q(\cdot,\rho) \in H^2_\delta(\mathbb{R}^n)$ *and for* $0 \le s \le 2$ *there exists* $C = C(n,s,\delta) > 0$ *such that*

$$||\psi_q(\cdot,\rho)||_{H^2} \le \frac{C}{|\rho|^{1-s}}. \tag{34}$$

Sketch of the proof of Theorem 3. Let u be a solution of (32) of type (33), then ψ_q must satisfy

$$(\nabla^2 + 2\rho\cdot\nabla - q)\psi_q = q. \tag{35}$$

The idea is that Eq. (35) can be solved for ψ_q by constructing an inverse for $(\nabla^2 + 2\rho\cdot\nabla)$ and solving the integral equation

$$\psi_q = (\nabla^2 + 2\rho\cdot\nabla)^{-1}\left(q(1+\psi_q)\right) \tag{36}$$

for ψ_q. For more details about how to solve the above equation, refer to [110, Lemma 5.2] where it is shown that the integral equation (36) can only be solved in $L^2_\delta(\mathbb{R}^n)$ for large $|\rho|$. □

Other approaches for the construction of CGO solutions for the Schrödinger equation have been considered in [36, 49]. The reader may refer to [110] for more details about references on this topic and a more in-depth explanation about the constructions of this kind of solutions.

Dirichlet-to-Neumann Map and Cauchy Data for the Schrödinger Equation

If 0 is not a Dirichlet eigenvalue for the Schrödinger equation, then the Dirichlet-to-Neumann map associated with a potential q can be defined by

$$\tilde{\Lambda}_q(f) = \frac{\partial u}{\partial \nu}|_{\partial\Omega},$$

where u solves the Dirichlet problem

$$\begin{cases} (\nabla^2 - q)u = 0 & \text{in } \Omega \\ u|_{\partial\Omega} = f. \end{cases}$$

As a consequence of Lemma 1, for any $q = \frac{\nabla^2\gamma^{1/2}}{\gamma^{1/2}}$,

$$\begin{aligned}\tilde{\Lambda}_q(f) &= \frac{\partial}{\partial \nu}(\gamma^{\frac{1}{2}}\gamma^{-\frac{1}{2}}\, u)|_{\partial\Omega} \\ &= \left(\frac{\partial \gamma^{\frac{1}{2}}}{\partial \nu}\,(\gamma^{-\frac{1}{2}}u) + \gamma^{\frac{1}{2}}\,\frac{\partial(\gamma^{-\frac{1}{2}}u)}{\partial \nu}\right)|_{\partial\Omega} \\ &= \left(\frac{1}{2}\gamma^{-\frac{1}{2}}\,\frac{\partial \gamma}{\partial \nu}\,\gamma^{-\frac{1}{2}} + \gamma^{\frac{1}{2}}\,\frac{\partial(\gamma^{-\frac{1}{2}}u)}{\partial \nu}\right)|_{\partial\Omega} \\ &= \frac{1}{2}\left(\gamma^{-1}\frac{\partial \gamma}{\partial \nu}\right)|_{\partial\Omega}\, f + \gamma^{\frac{1}{2}}|_{\partial\Omega}\,\Lambda_\gamma\big(\gamma^{-\frac{1}{2}}|_{\partial\Omega}\, f\big).\end{aligned}$$

So the two Dirichlet-to-Neumann maps $\tilde{\Lambda}_q$ and Λ_γ are related in the following way

$$\tilde{\Lambda}_q(f) = \frac{1}{2}\left(\gamma^{-1}\frac{\partial \gamma}{\partial \nu}\right)|_{\partial\Omega}\, f + \gamma^{\frac{1}{2}}|_{\partial\Omega}\,\Lambda_\gamma\big(\gamma^{-\frac{1}{2}}|_{\partial\Omega}\, f\big), \tag{37}$$

for any $f \in H^{\frac{1}{2}}(\partial\Omega)$. For $q \in L^\infty(\partial\Omega)$, the Cauchy data are defined as the set

$$\mathbf{C_q} = \left\{\left(u|_{\partial\Omega},\ \frac{\partial u}{\partial \nu}|_{\partial\Omega}\right) \mid u \in H^1(\Omega), \quad (\nabla^2 - q)u = 0 \quad \text{in} \quad \Omega\right\}.$$

If 0 is not an eigenvalue of $\nabla^2 - q$, then C_q is the graph given by

$$\mathbf{C_q} = \left\{\big(f, \tilde{\Lambda}_q(f)\big) \in H^{\frac{1}{2}}(\partial\Omega) \times H^{-\frac{1}{2}}(\partial\Omega)\right\}.$$

The result so far is very general and holds in any dimension $n \geq 2$. In the rest of the discussion on the uniqueness of Calderón's problem, a distinction is made between the higher dimensional case $n \geq 3$ and the two-dimensional one.

Global Uniqueness for $n \geq 3$

Sylvester and Uhlmann proved in [107] a result of global uniqueness for $C^2(\bar{\Omega})$ conductivities by solving in this way the identifiability question with the following result. Their result follows in dimension $n \geq 3$ from a more general one for the Schrödinger equation, which is useful in its own right for other inverse problems.

Theorem 4. *Let $q_i \in L^\infty(\Omega)$, i=1, 2. Assume $C_{q_1} = C_{q_2}$, then $q_1 = q_2$.*

Proof of Theorem 4. This result has been proved by constructing oscillatory solutions of $(\nabla^2 - q_i)\, u_i = 0$ in $\mathbb{R}^n$ with high frequencies. Beginning with the following equality

$$\int_\Omega (q_1 - q_2) u_1 u_2 = 0 \tag{38}$$

is true for any $u_i \in H^1(\Omega)$ solution to

$$(\nabla^2 - q_i)u_i = 0 \qquad \text{in} \quad \Omega, \qquad i = 1, 2.$$

Equality (38) follows by

$$\int_\Omega (q_1 - q_2)u_1 u_2 = \int_{\partial\Omega} \left(\frac{\partial u_1}{\partial \nu} u_2 - u_1 \frac{\partial u_2}{\partial \nu} \right) dS,$$

which can be easily obtained by the divergence theorem. This extends q_i on the whole $\mathbb{R}^n$ by taking $q_i = 0$ on $\mathbb{R}^n \setminus \Omega$ and taking solutions of

$$(\nabla^2 - q_i)u_i = 0 \qquad \text{in} \quad \mathbb{R}^n, \qquad i = 1, 2$$

of the form

$$u_i = \mathrm{e}^{x \cdot \rho_i} \left(1 + \psi_{q_i}(x, \rho_i)\right), \qquad i = 1, 2, \tag{39}$$

with $|\rho_i|$ large. This type of solutions are known as CGO solutions. $\rho_i, i = 1, 2$ is chosen of type

$$\begin{aligned} \rho_1 &= \frac{\eta}{2} + i \left(\frac{k + l}{2} \right) \\ \rho_2 &= -\frac{\eta}{2} + i \left(\frac{k - l}{2} \right), \end{aligned} \tag{40}$$

with $\eta, k, l \in \mathbb{R}^n$ and satisfying

$$\eta \cdot k = k \cdot k = \eta \cdot l = 0, \qquad |\eta|^2 = |k|^2 + |l|^2, \tag{41}$$

the choices of η, k, l having been made so that $\rho_i \cdot \rho_i = 0$, $i = 1, 2$. With these choices of $\rho_i, i = 1, 2$,

$$\begin{aligned} u_1 u_2 &= \left[\mathrm{e}^{x \cdot \frac{\eta}{2} + ix \cdot \left(\frac{k+l}{2} \right)} + \mathrm{e}^{x \cdot \frac{\eta}{2} + ix \cdot \left(\frac{k+l}{2} \right)} \psi_{q_1} \right] \\ &\quad \cdot \left[\mathrm{e}^{-x \cdot \frac{\eta}{2} + ix \cdot \left(\frac{k-l}{2} \right)} + \mathrm{e}^{-x \cdot \frac{\eta}{2} + ix \cdot \left(\frac{k-l}{2} \right)} \psi_{q_2} \right] \\ &= \mathrm{e}^{ix \cdot k} \left(1 + \psi_{q_1} + \psi_{q_2} + \psi_{q_1} \psi_{q_2}\right) \end{aligned}$$

and therefore

$$\widehat{(q_1 - q_2)}(-k) = -\int_\Omega \mathrm{e}^{ix \cdot k} (q_1 - q_2)(\psi_{q_1} + \psi_{q_2} + \psi_{q_1} \psi_{q_2}) dx. \tag{42}$$

By recalling that

$$||\psi_{q_i}||_{L^2(\Omega)} \leq \frac{C}{|\rho_i|}$$

and letting $|l| \to \infty$ one obtains $q_1 = q_2$ (see [110, proof of Theorem 6.2, pp 10]). □

As a consequence of this result, the result [107] stated below is finally obtained.

Theorem 5. *Let $\gamma_i \in C^2(\bar{\Omega})$, γ_i strictly positive, i=1, 2. If $\Lambda_{\gamma_1} = \Lambda_{\gamma_2}$, then $\gamma_1 = \gamma_2$ in $\bar{\Omega}$.*

Theorem 5 has been proved in [107] in a straightforward manner by constructing highly oscillatory solutions to $L_\gamma u = 0$ in Ω. In this chapter, the line of [110] is followed in the exposition of such result as a consequence of the more general Theorem 4. Such a choice has been made because of the clearer exposition made in [110].

One can proceed by showing that Theorem 4 implies Theorem 5 for the sake of completeness. The reader can find it also in [110]. The argument used is the following. Let $\gamma_i \in C^2(\bar{\Omega})$ be strictly positive and $\Lambda_{\gamma_1} = \Lambda_{\gamma_2}$. Then by [75],

$$\gamma_1|_{\partial\Omega} = \gamma_2|_{\partial\Omega},$$
$$\frac{\partial\gamma_1}{\partial\nu}|_{\partial\Omega} = \frac{\partial\gamma_2}{\partial\nu}|_{\partial\Omega},$$

therefore (37) implies $C_{q_1} = C_{q_2}$ i.e., $q_1 = q_2 =: q$ because of Theorem 4. Recall that

$$q_i = \frac{\nabla^2\gamma_i^{1/2}}{\gamma_i^{1/2}}, \qquad i = 1, 2,$$

which leads to

$$\nabla^2\gamma_1^{\frac{1}{2}} - q\,\gamma_1^{\frac{1}{2}} = 0$$
$$\nabla^2\gamma_2^{\frac{1}{2}} - q\,\gamma_2^{\frac{1}{2}} = 0$$

i.e.,

$$\nabla^2\left(\gamma_1^{\frac{1}{2}} - \gamma_2^{\frac{1}{2}}\right) - q\left(\gamma_1^{\frac{1}{2}} - \gamma_2^{\frac{1}{2}}\right) = 0$$

with

$$\left(\gamma_1^{\frac{1}{2}} - \gamma_2^{\frac{1}{2}}\right)|_{\partial\Omega} = 0.$$

Therefore it must be that

$$\gamma_1 = \gamma_2 \qquad \text{in} \quad \Omega,$$

by uniqueness of the solution of the Cauchy problem.

The identifiability question was then pushed forward to the case of $\gamma \in C^{1,1}(\bar{\Omega})$ with an affirmative answer by Nachman et al. in 1988 [90]. Nachman extended then this result to domains with $C^{1,1}$ boundaries (see [88]). The condition on the boundary was relaxed to $\partial\Omega$ Lipschitz by Alessandrini in 1990 in [3]; he proved uniqueness at the boundary and gave stability estimates for $\gamma \in W^{1,p}(\Omega)$, with $p > n$ by making use of singular solutions with an isolated singularity at the center of a ball. This method enables one to construct solutions of $L_\gamma u = 0$ on a ball behaving asymptotically like the singular solutions of the Laplace–Beltrami equation with separated variables. His results hold in dimension $n \geq 2$. Results of global uniqueness in the interior were also found in [3] among piecewise analytic perturbations of γ, giving an extension of Kohn and Vogelius result in [76] to Lipschitz domains.

Going back to the issue of global uniqueness, Brown [22] relaxed the regularity of the conductivity to $\frac{3}{2} + \varepsilon$ derivatives, which was followed by the uniqueness result of Päivärinta et al. [95] for $W^{\frac{3}{2},\infty}$ conductivities. Their result is based on the construction of CGO solutions for conductivities $\gamma \in W^{1,\infty}(\mathbb{R}^n)$ ($n \geq 2$). Recalling in what follows, their construction of the CGO followed by their uniqueness result.

Theorem 6 ([95]). *Let $\gamma \in W^{1,\infty}(\mathbb{R}^\kappa)$, γ strictly positive and $\gamma = 1$ outside a large ball. Let $-1 < \delta < 0$, then for $|\rho|$ sufficiently large there is a unique solution of*

$$\operatorname{div}(\gamma \nabla u) = 0 \quad \text{in} \quad \mathbb{R}^\kappa$$

of the form

$$u = \mathrm{e}^{x\cdot\rho}\left(\gamma^{-\frac{1}{2}} + \psi_\gamma(x,\rho)\right), \tag{43}$$

with $\psi_\gamma \in L^2_\delta(\mathbb{R}^n)$. Moreover, ψ_γ has the form

$$\psi_\gamma(x,\rho) = \left(\omega_0(x,\rho) - \gamma^{-\frac{1}{2}}\right) + \omega_1(x,\rho), \tag{44}$$

where ω_0, ω_1 satisfy

$$\lim_{|\rho|\to\infty} ||\omega_0(x,\rho) - \gamma^{-\frac{1}{2}}||_{H^1_\delta} = 0 \tag{45}$$

and

$$\lim_{|\rho|\to\infty} ||\omega_1(x,\rho)||_{L^2_\delta} = 0. \tag{46}$$

Here, the idea behind the proof of the above Theorem 6 is recalled. The first step is to rewrite the conductivity equation

$$\operatorname{div}(\gamma\nabla u) = 0 \quad \text{in} \quad \mathbb{R}^n$$

as

$$\Delta u + A\cdot\nabla u = 0 \quad \text{in} \quad \mathbb{R}^n,$$

where

$$A + \nabla\log\gamma \in L^\infty(\mathbb{R}^n)$$

has compact support. By introducing

$$\Phi_\varepsilon = \varepsilon^{-n}\left(\frac{x}{\varepsilon}\right),$$

with $\Phi(x)$ a mollifier, one can define

$$\begin{aligned}\varphi_\varepsilon &= \Phi_\varepsilon * \log\gamma\\ A_\varepsilon &= \Phi_\varepsilon * A\\ \omega_0(x,\varepsilon) &= \mathrm{e}^{-\frac{\varphi_\varepsilon(x)}{2}}\end{aligned}$$

and with the above choice of ω_0 one can show that

$$\lim_{\varepsilon\to 0} ||\,\omega_0 - \gamma^{-\frac{1}{2}}\,||_{H^1_\delta} = 0. \tag{47}$$

Let $\rho\in\mathbb{C}^n$ be such that $\rho\cdot\rho = 0$ and define the operators

$$\Delta_\rho u := \mathrm{e}^{-x\cdot\rho}\Delta(\mathrm{e}^{x\cdot\rho}u) = \Delta u + 2\rho\cdot\nabla u \tag{48}$$

$$\nabla_\rho u := \mathrm{e}^{-x\cdot\rho}\nabla(\mathrm{e}^{x\cdot\rho}u) = \nabla u + \rho u. \tag{49}$$

One can then define for any $f\in C_0^\infty(\mathbb{R}^n)$

$$\Delta_\rho^{-1} f = \frac{1}{(2\pi)^n}\int \mathrm{e}^{ix\cdot\xi}\frac{\hat f(\xi)}{-|\xi|^2+2i\rho\cdot\xi}\,d\xi,$$

which can then be extended to a bounded operator

$$\Delta_\rho^{-1} : H^s_{\delta+1}(\mathbb{R}^n) \to H^s_\delta(\mathbb{R}^n),$$

for any $-1 < \delta < 0$ and $s \geq 0$. Moreover

$$||\Delta_\rho^{-1}||_{H^s_{\delta+1} \to H^s_\delta} \leq \frac{C(s,\delta,n)}{|\rho|},$$

for some $C > 0$. The idea is now to construct ω_1 solution to

$$(\Delta_\rho + A \cdot \nabla_\rho)\omega_1 = -(\Delta_\rho + A \cdot \nabla_\rho)\omega_0, \tag{50}$$

recalling that $\omega_0(x,\varepsilon) = \mathrm{e}^{-\frac{\varphi_\varepsilon(x)}{2}}$ and therefore depends on ε. If one sets now

$$\omega_1 = \Delta_\rho^{-1}\tilde{\omega},$$

one can then rewrite (50) as

$$\underbrace{(I + A \cdot \nabla_\rho \Delta_\rho^{-1})}_{:=T_\rho(\gamma)}\,\tilde{\omega} = f_\varepsilon, \tag{51}$$

where

$$\begin{aligned} f_\varepsilon &:= -(\Delta_\rho + A \cdot \nabla_\rho)\omega_0 \\ &= -\mathrm{e}^{-\frac{\varphi_\varepsilon}{2}}\left(-\frac{1}{2}\Delta\varphi_\varepsilon + \frac{1}{4}(\nabla\varphi_\varepsilon)^2 - \frac{1}{2}A \cdot \nabla\varphi_\varepsilon + (A - A_\varepsilon)\cdot\rho\right). \end{aligned}$$

An approximate inverse of $T_\rho(\gamma)$ is given by

$$\begin{aligned} S_\rho(\gamma) &:= \gamma^{-\frac{1}{2}}(I - A \cdot \nabla_\rho \delta_\rho^{-1})\gamma^{\frac{1}{2}} \\ &= \gamma^{-\frac{1}{2}} T_\rho(\gamma^{-1})\gamma^{\frac{1}{2}}, \end{aligned}$$

therefore (51) has a unique solution in an appropriate space. To study now the behavior of ω_1 as $\varepsilon \to 0$ and $|\rho| \to \infty$, recall that

$$\omega_1(x,\varepsilon,\rho) = \Delta_\rho^{-1} S_\rho f_\varepsilon + \underbrace{\Delta_\rho^{-1}(T_\rho^{-1} - S_\rho) f_\varepsilon}_{:=h_\rho} \tag{52}$$

and now one can show that

$$\lim_{|\rho|\to\infty} ||h_\rho||_{H^1_\delta(\mathbb{R}^n)} = 0, \tag{53}$$

which concludes the proof. Theorem 43 is then used in [95] to prove global uniqueness for conductivities $\gamma \in W^{\frac{3}{2},\infty}(\Omega)$, where Ω is a bounded domain in $\mathbb{R}^n$, with $n \geq 3$. More precisely they prove the following theorem.

Theorem 7. *Let $n \geq 3$. Let $\gamma_i \in W^{\frac{3}{2},\infty}(\Omega)$ be strictly positive on $\bar{\Omega}$, $i = 1, 2$. If*

$$\Lambda_{\gamma_1} = \Lambda_{\gamma_2},$$

then

$$\gamma_1 = \gamma_2 \quad on \quad \Omega.$$

Again, the idea is to give to the reader a flavor of how Theorem 7 is proven in [95]. The main idea is that if $\gamma_i \in W^{1,\infty}(\Omega)$ and $\Lambda_{\gamma_1} = \Lambda_{\gamma_2}$, with $a_i = \sqrt{\gamma_i}$, $i = 1, 2$, then the identity

$$\int_\Omega (\nabla a_1 \cdot \nabla(u_1 v) - \nabla a_2 \cdot \nabla(u_2 v)) - \int_\Omega (\nabla(a_1 u_1) \cdot \nabla v - \nabla(a_2 u_2) \cdot \nabla v)\ dx = 0, \tag{54}$$

holds true $\forall v \in H^1(\Omega)$, $\forall u_i \in H^1(\Omega)$ solution to $\mathrm{div}(\gamma_i \nabla u_i) = 0$ in $\Omega = 1, 2$. The reader should notice that so far the results obtained hold true for conductivities of type $\gamma \in W^{1,\infty}(\Omega)$. It is at this stage that one needs to assume

$$\gamma \in W^{\frac{3}{2},\infty}(\mathbb{R}^n)$$

to show the following technical results for $\omega_1(x, \rho) = \omega_1(x, \varepsilon, \rho)$ as in (52), $\varepsilon = |\rho|^{-1}$

$$\lim_{|\rho| \to \infty} \int e^{ix \cdot \xi} \nabla \gamma^{\frac{1}{2}} \cdot \nabla \omega_1\ dx = 0. \tag{55}$$

With this choice of ω_1, by substituting the CGO solutions (44) into identity (54), one then can gain the desired uniqueness result.

The above result was then followed by uniqueness for $W^{\frac{3}{2},p}$ (with $p > 2n$) in [23]. Recently Haberman and Tataru [48] proved that uniqueness holds for C^1 conductivities and Lipschitz conductivities close to the identity. Their result is the following.

Theorem 8. *Let $n \geq 3$ and $\Omega \subset \mathbb{R}^n$ be a bounded domain with Lipschitz boundary. Let $\gamma_i \in W^{1,\infty}(\bar{\Omega})$ be a real conductivity, $i = 1, 2$. Suppose there exists a constant $C = C(n, \Omega)$ such that γ_i, $i = 1, 2$ satisfies either*

$$||\nabla \log \gamma_i||_{L^\infty(\bar{\Omega})} \leq C \tag{56}$$

or

$$\gamma_i \in C^1(\bar{\Omega}). \tag{57}$$

If

$$\Lambda_{\gamma_1} = \Lambda_{\gamma_2},$$

then

$$\gamma_1 = \gamma_2.$$

Global Uniqueness in the Two-Dimensional Case

The two-dimensional inverse conductivity problem must often be treated as a special case. Although results in [76] gave a positive answer to the identifiability question in the case of piecewise analytic conductivities, it was not until 1996 that Nachman [89] proved a global uniqueness result to Calderón problem for conductivities in $W^{2,p}(\Omega)$, for some $p > 1$. An essential part of his argument is based on the construction of the CGO solutions and the $\bar{\partial}$-method (sometimes written "d-bar method") in inverse scattering introduced in one dimension by Beals and Coifman (see [14,15]). The result of [89] has been improved in 1997 for conductivities having one derivative in an appropriate sense (see [24]) and the question of uniqueness was settled in $L^\infty(\Omega)$ finally by Astala and Päivärinta [10] using $\bar{\partial}$-methods. They proved

Theorem 9. *Let Ω be a bounded domain in $\mathbb{R}^2$ and $\gamma_i \in L^\infty$, $i = 1, 2$ be real functions such that for some constant M, $M^{-1} < \gamma_i < M$. Then*

$$\Lambda_{\gamma_1} = \Lambda_{\gamma_2} \implies \gamma_1 = \gamma_2.$$

Let us first explain the complex version of the problem used by [10]. Using the complex variable $z = x_1 + \mathrm{i}x_2$, and the notation $\partial = \partial/\partial z, \bar{\partial} = \partial/\partial\bar{z}$. Then the following result is available [10]:

Lemma 2. *Let Ω be the unit disk in the plane and $u \in H^1(\Omega)$ be a solution of $L_\gamma u = 0$. Then there exists a real function $v \in H^1(\Omega)$, unique up to a constant, such that $f = u + \mathrm{i}v$ satisfies the Beltrami equation*

$$\bar{\partial} f = \mu \overline{\partial f}, \tag{58}$$

where $\mu = (1 - \gamma)/(1 + \gamma)$.

Conversely if $f \in H^1(\Omega)$ satisfies (58), with a real valued μ, then $u = \mathrm{Re}\, f$ and $v = \mathrm{Im}\, f$ satisfy

$$L_\gamma u = 0 \quad \text{and} \quad L_{\gamma^{-1}} v = 0, \tag{59}$$

where $\gamma = (1 - \mu)/(1 + \mu)$.

Astala and Päivärinta reduce the general case of Ω to that of the disk, and show that the generalized Hilbert transform $\mathcal{H}_\mu : u|_{\partial\Omega} \mapsto v|_{\partial\Omega}$ uniquely determines, and is determined by Λ_γ. They go on to construct CGO solutions to (58) of the form

$$f_\mu(z,k) = \mathrm{e}^{ikz}\left(1 + O\left(\frac{1}{z}\right)\right) \text{ as } |z| \to \infty \tag{60}$$

and using a result connecting pseudoanalytic functions with quasi-regular maps prove that $\mathcal{H}_\mu$ determines μ. The original method Nachman used to prove uniqueness has resulted in the development of $\bar{\partial}$ reconstruction methods which are described below (section "Direct Nonlinear Solution"). See also the work of Druskin [38], which provides some answers to the 2-D geophysical settings.

Some Open Problems for the Uniqueness

One of the main open problems in dimension $n \geq 3$ is to investigate whether global uniqueness holds for the minimal assumption $\gamma \in L^\infty(\Omega)$ or else to find what are the minimal assumptions on γ in order to guarantee uniqueness. The inverse conductivity problem makes sense for conductivities that are indeed merely L^∞. There are neither proofs nor counter-examples for this in any dimension, to the authors' knowledge, but it has been conjectured by Uhlmann that the optimal assumption is that the conductivities are Lipschitz. These open problems influence of course also the stability issue of finding appropriate assumptions (possibly on γ) in order to improve the unstable nature of EIT. This issue will be studied in the next section.

Stability of the Solution at the Boundary

The result of uniqueness at the boundary of Theorem 2 has been improved in [108] to a stability estimate. The result is the following.

Theorem 10. *Let $\gamma_i \in C^\infty(\bar{\Omega})$, $i = 1, 2$, satisfy*

$$0 < \frac{1}{E} \leq \gamma_i \leq E, \qquad i = 1, 2 \tag{61}$$

$$||\gamma_i||_{C^2(\bar{\Omega})} \leq E, \qquad i = 1, 2. \tag{62}$$

Given any $0 < \sigma < \frac{1}{n+1}$, there exists $C = C(\Omega, E, n, \sigma)$ such that

$$||\gamma_1 - \gamma_2||_{L^\infty(\partial\Omega)} \leq C\,||\Lambda_{\gamma_1} - \Lambda_{\gamma_2}||_* \tag{63}$$

and

$$\left|\left|\frac{\partial\gamma_1}{\partial\nu} - \frac{\partial\gamma_2}{\partial\nu}\right|\right|_{L^\infty(\partial\Omega)} \leq C\,||\Lambda_{\gamma_1} - \Lambda_{\gamma_2}||_*^\sigma, \tag{64}$$

where $||\cdot||_*$ *denotes the norm in the space of bounded linear operators from* $H^{\frac{1}{2}}(\partial\Omega)$ *to* $H^{-\frac{1}{2}}(\partial\Omega)$.

This result improves the one of Theorem 2 in the sense that it is no longer necessary to require $\gamma \in C^\infty(\bar{\Omega})$ to determine γ itself and its derivative at the boundary. It is only needed that γ be continuous on $\bar{\Omega}$ to determine the boundary values of γ, where if $\gamma \in C^1(\bar{\Omega})$ then one can determine γ and its first derivative on $\partial\Omega$ as well. Subsequent results of stability at the boundary along the same lines have been proved in [3,7,22,43,88,91].

Global Stability for $n \geq 3$

In 1988 Alessandrini [2] proved that, in dimension $n \geq 3$, under an a priori assumption on γ of type

$$||\gamma||_{H^s(\Omega)} \leq E, \qquad \text{for some } s > \frac{n}{2} + 2,$$

γ depends continuously on Λ_γ with a modulus of continuity of logarithmic type. The result is stated below.

Theorem 11. *Let* $n \geq 3$. *Suppose that* $s > \frac{n}{2}$ *and that* $\gamma_i \in C^\infty(\bar{\Omega})$, $i = 1, 2$ *is a conductivity satisfying*

$$0 < \frac{1}{E} \leq \gamma_i \leq E, \qquad i = 1, 2 \tag{65}$$

$$||\gamma_i||_{H^{s+2}(\Omega)} \leq E, \qquad i = 1, 2. \tag{66}$$

Then there exists $C = C(\Omega, E, n, s)$ *and* $\tau = \tau(n, s)$, *with* $0 < \tau < 1$ *such that*

$$||\gamma_1 - \gamma_2||_{L^\infty(\Omega)} \leq C \left(|\log ||\Lambda_{\gamma_1} - \Lambda_{\gamma_2}||_*|^{-\tau} + ||\Lambda_{\gamma_1} - \Lambda_{\gamma_2}||_* \right). \tag{67}$$

It has been proved [3,4] that a similar stability estimate holds if (66) is replaced by

$$||\gamma_i||_{W^{2,\infty}(\Omega)} \leq E, \qquad i = 1, 2. \tag{68}$$

Mandache [86] proved that logarithmic stability is optimal for dimension $n \geq 2$ if the a priori assumption is of the form

$$||\gamma||_{C^k(\bar{\Omega})} \leq E, \tag{69}$$

for any finite $k = 0, 1, 2, \ldots$. One of the main open problems in the stability issue is then to improve this logarithmic-type stability estimate under some additional a priori condition. In [8] it has been shown that (67) can be improved to a Lipschitz-type estimate in the case in which γ is piecewise constant with jumps on a finite

number of domains. For piecewise constant complex conductivities a similar result has been proved in [16], where piecewise constant potentials of the Schrödinger equation have been investigated in [18] and Lipschitz stability estimates have been proved in this case as well. For more in-depth discussion about the stability in EIT and open problems in that regard refer to [5]. A similar estimate to (67) for the potential case can be found in [110].

Global Stability for the Two-Dimensional Case

Logarithmic-type stability estimates in dimension $n = 2$ were obtained by [12, 13, 82]. The results obtained in [82] last require only γ be Hölder continuous of positive exponent

$$||\gamma||_{C^{\alpha}(\bar{\Omega})} \leq E, \tag{70}$$

for some α, $0 < \alpha \leq 1$.

Some Open Problems for the Stability

The main open problem is to improve the logarithmic-type estimate found in [2] in any dimension $n \geq 2$. One approach would be to investigate whether the a priori regularity assumptions (68) can be further relaxed. On the other hand, since it has been observed [86] that this logarithmic type of estimate cannot be avoided under any a priori assumption of type (69) for any finite $k = 0, 1, 2, \ldots$, it seems natural to think that another direction to proceed would be the one of looking for different a priori assumptions rather than the one of type (69). For a complete analysis of open problems in this area, refer to [5].

The Anisotropic Case

The Non-uniqueness

In anisotropic media the conductivity depends on the direction, therefore it is represented by a matrix $\gamma = (\gamma_{ij})_{i,\,j=1}^{n}$, which is symmetric and positive definite. Anisotropic conductivity appears in nature, for example as a homogenization limit in layered or fibrous structures such as rock stratum or muscle, as a result of crystalline structure or of deformation of an isotropic material. Let $\Omega \subset \mathbb{R}^n$ be a domain with smooth boundary $\partial\Omega$ (a Lipschitz boundary will be enough in most cases). The Dirichlet problem associated in the anisotropic case takes the form

$$\begin{cases} \sum\limits_{i,j=1}^{n} \frac{\partial}{\partial x^i}(\gamma_{ij}\ \frac{\partial u}{\partial x^j}) = 0 \text{ in} \quad \Omega \\ u|_{\partial\Omega} = f, \end{cases} \tag{71}$$

where $f \in H^{\frac{1}{2}}(\partial\Omega)$ is a prescribed potential at the boundary. The Dirichlet-to-Neumann map associated with γ is defined by

$$\Lambda_\gamma f = \gamma \nabla u \cdot \nu|_{\partial\Omega}, \tag{72}$$

for any u solution to (71). Here $\gamma \nabla u \cdot \nu = \sum_{i,j=1}^{n} \left(\gamma_{ij} \frac{\partial u}{\partial x^j}\right) \nu_i|_{\partial\Omega}$ and as usual $\nu = (\nu_i)_{i=1}^n$ is the unit outer normal to $\partial\Omega$. The weak formulation of (72) is commonly used and will be given below for the sake of completeness.

Definition 1. The Dirichlet-to-Neumann map associated with (71) is

$$\Lambda_\gamma : H^{\frac{1}{2}}(\partial\Omega) \longrightarrow H^{-\frac{1}{2}}(\partial\Omega)$$

given by

$$\langle \Lambda_\gamma f, \eta \rangle = \int_\Omega \sigma(x) \nabla u(x) \cdot \nabla \phi(x)\, dx, \tag{73}$$

for any $f, \eta \in H^{\frac{1}{2}}(\partial\Omega)$, $u, \phi \in H^1(\Omega)$, $\phi|_{\partial\Omega} = \eta$ and u is the weak solution to (71).

A conductor is isotropic when $\gamma = (\gamma_{ij})$ is rotation invariant, i.e., when at each point

$$R^T \gamma R = \gamma,$$

for all rotations R. This is the case exactly when $\gamma = \alpha\, I$, where $\alpha > 0$ is a scalar function and I the identity matrix.

In the section "The Isotropic Case," the uniqueness problem for the isotropic case was considered solved; on the other hand, in the anisotropic case, Λ_γ does not in general determine γ. Tartar (see [75]) observed the following non-uniqueness result.

Proposition 1. *If $\psi : \bar{\Omega} \longrightarrow \bar{\Omega}$ is a C^1 diffeomorphism such that $\psi(x) = x$, for each $x \in \partial\Omega$, then γ and $\tilde{\gamma} = \frac{(D\psi)\gamma(D\gamma)^T}{\det(D\psi)} \circ \psi^{-1}$ have the same Dirichlet-to-Neumann map.*

The proof of this result is given below as a tutorial for the first-time reader of this material.

Proof. Let us consider the change of variables $y = \psi(x)$ on the Dirichlet integral

$$\int_\Omega \gamma_{ij}(x) \frac{\partial u}{\partial x^i} \frac{\partial u}{\partial x^j}\, dx = \int_\Omega \tilde{\gamma}_{ij}(y) \frac{\partial \tilde{u}}{\partial y^i} \frac{\partial \tilde{u}}{\partial y^j}\, dx \tag{74}$$

where

$$\tilde{\gamma}(y) = \frac{(D\psi)\,\gamma(D\psi)^T}{\det(D\psi)} \circ \psi^{-1}(y)$$

and

$$\tilde{u}(y) = u \circ \psi^{-1}(y).$$

Notice that the solution u of the Dirichlet problem

$$\begin{cases} \nabla \cdot \gamma \nabla u = 0 \text{ in } \quad \Omega \\ u|_{\partial\Omega} = f \end{cases}$$

minimizes the integral appearing on the left-hand side of (74), therefore $\tilde{u} = u \circ \psi^{-1}$ minimizes the Dirichlet integral appearing on the right-hand side of the same. One can then conclude that $\tilde{u}$ solves

$$\begin{cases} \nabla \cdot (\tilde{\gamma} \nabla \tilde{u}) = 0 & \text{in } \quad \Omega \\ \tilde{u}|_{\partial\Omega} = \tilde{f} = u \circ \psi^{-1}. \end{cases}$$

Let us consider now the solution v of

$$\begin{cases} \nabla \cdot (\gamma \nabla v) = 0 \text{ in } \quad \Omega \\ v|_{\partial\Omega} = g \end{cases}$$

and let $\tilde{v}$ be obtained by v by the change of variable, therefore $\tilde{v}$ solves

$$\begin{cases} \nabla \cdot (\tilde{\gamma} \nabla \tilde{v}) = 0 & \text{in } \quad \Omega \\ \tilde{v}|_{\partial\Omega} = \tilde{g} = g \circ \psi^{-1}. \end{cases}$$

By the change of variables in the Dirichlet integrals,

$$\int_\Omega \gamma_{ij} \frac{\partial u}{\partial x^i} \frac{\partial v}{\partial x^j}\, dx = \int_\Omega \tilde{\gamma}_{ij} \frac{\partial \tilde{u}}{\partial y^i} \frac{\partial \tilde{v}}{\partial y^j}\, dy,$$

which can be written as

$$\int_\Omega \gamma \nabla u \cdot \nabla v\, dx = \int_\Omega \tilde{\gamma} \nabla \tilde{u} \cdot \nabla \tilde{v}\, dy,$$

which is equivalent to

$$\int_\Omega \nabla \cdot (v\gamma \nabla u)\, dx - \int_\Omega v \nabla \cdot (\gamma \nabla u)\, dx = \int_\Omega \nabla \cdot (\tilde{v}\tilde{\gamma} \nabla \tilde{u})\, dy - \int_\Omega \tilde{v} \nabla \cdot (\tilde{\gamma} \nabla \tilde{u})\, dy$$

and by the divergence theorem

$$\int_{\partial\Omega} v\gamma\nabla u\cdot\nu\, ds = \int_{\partial\Omega} \tilde{v}\tilde{\gamma}\nabla\tilde{u}\cdot\nu\, ds,$$

but $\tilde{v} = v\circ\psi^{-1} = v = g$ and $\tilde{u} = u\circ\psi^{-1} = u = f$ at the boundary $\partial\Omega$, then

$$\int_{\partial\Omega} g\Lambda_\gamma(f)\, ds = \int_{\partial\Omega} g\Lambda_{\tilde{\gamma}}(f)\, ds$$

then $\Lambda_\gamma = \Lambda_{\tilde{\gamma}}$. □

Since Tartar's observation has been made, different lines of research have been pursued. One direction was to prove the uniqueness of γ up to diffeomorphisms that fix the boundary, whereas the other direction was to study conductivities with some a priori information. The first direction of research is summarized in what follows.

Uniqueness up to Diffeomorphism

The question here is to investigate whether Tartar's observation is the only obstruction to unique identifiability of the conductivity. A first observation is that the physical problem of determining the conductivity of a body is closely related to the geometrical problem of determining a Riemannian metric from its Dirichlet-to-Neumann map for harmonics functions [80].

Let (M, g) be a compact Riemannian manifold with boundary. The Laplace–Beltrami operator associated with the metric g is given in local coordinates by

$$\Delta_g := \sum_{i\ j=1}^{n} (\det g)^{-\frac{1}{2}} \frac{\partial}{\partial x^i}\left\{(\det g)^{\frac{1}{2}}\, g^{ij}\, \frac{\partial u}{\partial x^j}\right\}.$$

The Dirichlet-to-Neumann map associated with g is the operator Λ_g mapping functions $u|_{\partial M} \in H^{1/2}(\partial M)$ into $(n-1)$-forms $\Lambda_\sigma(u|_{\partial M}) \in H^{-1/2}\left(\Omega^{n-1}(\partial M)\right)$

$$\Lambda_g(f) = i^\star(*_g du), \tag{75}$$

for any $u \in H^1(M)$ solution to $\Delta_g u = 0$ in M, with $u|_{\partial M} = f$. Here i is the inclusion map $i : \partial M \to M$ and $i^\star$ denotes the pull-back of forms under the map i. In any local coordinates (75) becomes

$$\Lambda_g(f) = \sum_{i,j=1}^{n} \nu^i g^{ij} \frac{\partial u}{\partial x_j}\sqrt{\det g}|_{\partial M}. \tag{76}$$

The inverse problem is to recover g from Λ_g. In dimension $n \geq 3$, the conductivity γ uniquely determines a Riemannian metric g such that

$$\gamma = *_g, \tag{77}$$

where $*_g$ is the Hodge operator associated the metric g mapping 1-forms on M into $(n-1)$-forms (see [42, 80, 81]). In any local coordinates (77) becomes

$$(g_{ij}) = (\det \gamma_{kl})^{\frac{1}{n-2}}(\gamma_{ij}) \quad \text{and} \quad (\gamma_{ij}) = (\det g_{kl})^{\frac{1}{2}}(g^{ij}), \tag{78}$$

where (g^{ij}), (γ^{ij}) denotes the matrix inverse of (g_{ij}) and (γ_{ij}) respectively. It has been shown in [80] that if M is a domain in $\mathbb{R}^n$, then for $n \geq 3$

$$\Lambda_g = \Lambda_\gamma. \tag{79}$$

In dimension $n \geq 3$ if ψ is a diffeomorphism of $\bar{M}$ that fixes the boundary,

$$\Lambda_{\psi^* g} = \Lambda_g, \tag{80}$$

where $\psi^* g$ is the pull-back of g under ψ. For the case $n = 2$ the situation is different as the two-dimensional conductivity determines a conformal structure of metrics under scalar field, i.e., there exists a metric g such that $\gamma = \varphi *_g$, for a positive function φ. Therefore in $n = 2$, if ψ is a diffeomorphism of $\bar{M}$ that fixes the boundary,

$$\Lambda_{\varphi\psi^* g} = \Lambda_g, \tag{81}$$

for any smooth positive function such that $\varphi|_{\partial M} = 1$. It seems natural to think that (80) and (81) are the only obstructions to uniqueness for $n \geq 3$ and $n = 2$ respectively. In 1989 Lee and Uhlmann [80] formulated the following two conjectures.

Conjecture 1. Let $\bar{M}$ be a smooth, compact n-manifold, with boundary, $n \geq 3$ and let g, $\tilde{g}$ be smooth Riemannian metrics on $\bar{M}$ such that

$$\Lambda_g = \Lambda_{\tilde{g}}.$$

Then there exists a diffeomorphism $\psi : \bar{M} \to \bar{M}$ with $\psi|_{\partial M} = Id$, such that $g = \psi^\star \tilde{g}$.

Conjecture 2. Let $\bar{M}$ be a smooth, compact 2-manifold with boundary, and let g, $\tilde{g}$ be smooth Riemannian metrics on $\bar{M}$ such that

$$\Lambda_g = \Lambda_{\tilde{g}}.$$

Then there exists a diffeomorphism $\psi : \bar{M} \to \bar{M}$ with $\psi|_{\partial M} = Id$, such that $\psi^\star\, \tilde{g}$ is a conformal multiple of g, in other words there exists $\phi \in C^\infty(\bar{M})$ such that

$$\psi^\star\, \tilde{g} = \phi\, g.$$

Conjecture 1 has been proved in [80] in a particular case. The result is the following.

Theorem 12. *Let $\bar{M}$ be a compact, connected, real-analytic n-manifold with connected real-analytic boundary, and assume that $\pi_1(\bar{M}, \partial M) = 0$ (this assumption means that every closed path in $\bar{M}$ with base point in ∂M is homotopic to some path that lies entirely in ∂M). Let g and $\tilde{g}$ be real-analytic metrics on $\bar{M}$ such that*

$$\Lambda_g = \Lambda_{\tilde{g}},$$

and assume that one of the following conditions holds:

1. *$\bar{M}$ is strongly convex with respect to both g and $\tilde{g}$;*
2. *either g or $\tilde{g}$ extends to a complete real-analytic metric on a non-compact real-analytic manifold $\tilde{M}$ (without boundary) containing $\bar{M}$.*

Then there exists a real-analytic diffeomorphism $\psi : \bar{M} \to \bar{M}$ with $\psi|_{\partial M} = Id$, such that $g = \psi^\star \tilde{g}$.

Theorem 12 has been proved by showing that one can recover the full Taylor series of the metric at the boundary from Λ_g. The diffeomorphism ψ is then constructed by analytic continuation from the boundary. As previously mentioned, the full Taylor series of γ was recovered by Kohn and Vogelius in [75] from the knowledge of Λ_γ in the isotropic case and then a new proof was given in [106] by showing that the full symbol of the pseudodifferential operator Λ_γ determines the full Taylor series of γ at the boundary. In [80] a simpler method suggested by R. Melrose consisting of factorizing Δ_g, is used. In 1990 Sylvester proved in [105] Conjecture 2 in a particular case. His result is the following.

Theorem 13. *Let Ω be a bounded domain in $\mathbb{R}^2$ with a C^3 boundary and let γ_1, γ_2 be anisotropic C^3 conductivities in $\bar{\Omega}$ such that*

$$\| \log(\det \gamma_i) \|_{C^3} < \varepsilon(M, \Omega), \quad for \quad i = 1, 2, \tag{82}$$

with $M \geq \| \gamma_i \|_{C^3}$, for i=1, 2 and $\varepsilon(M, \Omega)$ sufficiently small. If

$$\Lambda_{\gamma_1} = \Lambda_{\gamma_2},$$

then there exists a C^3 diffeomorphism ψ of $\bar{\Omega}$ such that $\psi|_{\partial\Omega} = Id$ and such that

$$\psi_\star \gamma_1 = \gamma_2.$$

Nachman [89] extended this result in 1995 by proving the same theorem but removing the hypothesis (82). In 1999 Lassas and Uhlmann [78] extended the result of [80]. They assumed that the Dirichlet-to-Neumann map is measured only on

a part of the boundary which is assumed to be real-analytic in the case $n \geq 3$ and C^∞-smooth in the two-dimensional case. The metric is here recovered (up to diffeomorphism) and the manifold is reconstructed. Since a manifold is a collection of coordinate patches, the idea is to construct a representative of an equivalent class of the set of isometric Riemannian manifolds $(M,\ g)$. Recalling that if Γ is an open subset of ∂M, and defining

$$\Lambda_{g,\Gamma}(f) = \Lambda_g(f)|_\Gamma,$$

for any f with $\operatorname{supp} f \subseteq \Gamma$. The main result of [78] is given below.

Theorem 14. *Let us assume that one of the following conditions is satisfied:*

1. *M is a connected Riemannian surface;*
2. *$n \geq 3$ and $(M,\ g)$ is a connected real-analytic Riemannian manifold and the boundary ∂M is real-analytic in the nonempty set $\Gamma \subset \partial M$.*

Then

1. *For $\dim M = 2$ the $\Lambda_{g,\Gamma}$-mapping and Γ determine the conformal class of the Riemannian manifold $(M,\ g)$.*
2. *For a real-analytic Riemannian manifold $(M,\ g)$, $\dim M > 2$ which boundary is real analytic in Γ, the $\Lambda_{g,\Gamma}$-mapping and Γ determine the Riemannian manifold $(M,\ g)$.*

This result improved the one in [80] also because here the only assumption on the topology of the manifold is the connectedness, while in [80] the manifold was simply connected and the boundary of the manifold was assumed to be geodesically convex. Theorem 14 has been extended in [79] to a completeness hypothesis on $\bar{M}$.

Anisotropy Which Is Partially A Priori Known

Another approach to the anisotropic problem is to assume that the conductivity γ is a priori known to depend on a restricted number of unknown spatially dependent parameters. In 1984 Kohn and Vogelius (see [75]) considered the case where the conductivity matrix $\gamma = (\gamma_{ij})$ is completely known with the exception of one eigenvalue. The main result is the following.

Theorem 15. *Let $\gamma,\ \tilde{\gamma}$ be two symmetric, positive definite matrices with entries in $L^\infty(\Omega)$, and let $\{\gamma_i\},\ \{\tilde{\gamma}_i\}$ and $\{e_i\},\ \{\tilde{e}_i\}$ be the corresponding eigenvalues and eigenvectors. For $x_0 \in \partial\Omega$, let B be a neighborhood of x_0 relative to $\bar{\Omega}$, and suppose that*

$$\gamma,\ \tilde{\gamma} \in C^\infty(B); \tag{83}$$

$$\partial\Omega \cap B \quad is \quad C^\infty; \tag{84}$$

$$e_j = \tilde{e}_j, \qquad \lambda_j = \tilde{\lambda}_j \quad in \quad B, \quad for \quad 1 \le j \le n-1; \tag{85}$$

$$e_n(x_0) \cdot \nu(x_0) \neq 0. \tag{86}$$

If

$$Q_\gamma(\phi) = Q_{\tilde{\gamma}}(\phi) \quad for\ every \quad \phi \in H^{\frac{1}{2}}(\partial\Omega),$$

with $\operatorname{supp}\phi \subset B \cap \partial\Omega$, *then*

$$D^k \tilde{\lambda}_n(x_0) = D^k \lambda_n(x_0),$$

for every $k = (k_1, \ldots, k_n)$, $k_i \in \mathbb{Z}^+$, $i = 1, \ldots, n$.

In 1990 Alessandrini [3] considered the case in which γ is a priori known to be of type

$$\gamma(x) = A(a(x)),$$

where $t \to A(t)$ is a given matrix-valued function and $a = a(x)$ is an unknown scalar function. He proved results of uniqueness and stability at the boundary and then uniqueness in the interior among the class of piecewise real-analytic perturbations of the parameter $a(x)$. The main hypothesis he used is the so-called monotonicity assumption

$$D_t A(t) \ge C\ I,$$

where $C > 0$ is a constant. In 1997 Lionheart [81] proved that the parameter $a(x)$ can be uniquely recovered for a conductivity γ of type

$$\gamma(x) = a(x)\ A_0(x),$$

where $A_0(x)$ is given. Results in [3] have been extended in 2001 by Alessandrini and Gaburro [6] to a class of conductivities

$$\gamma(x) = A(x,\ a(x)),$$

where $A(x,\ t)$ is given and satisfies the monotonicity condition with respect to the parameter t

$$D_t A(x,\ t) \ge C\ I,$$

where $C > 0$ is a constant (see [6] or [41] for this argument). In [6] the authors improved results of [3] since they relaxed the hypothesis on $A(\cdot,\ t)$ for the global uniqueness in the interior and the result there obtained can be applied to [81] as well.

The technique of [6] can also be applied to the so-called one-eigenvalue problem introduced in [75]. Results of [6] have been recently extended to manifolds [43] and to the case when the local Dirichlet-to-Neumann map is prescribed on an open portion of the boundary [7].

Some Remarks on the Dirichlet-to-Neumann Map

EIT with Partial Data

In many applications of EIT one can actually only take measurements of voltages and currents on some portion of the boundary. In such situation the Dirichlet-to-Neumann map can only be defined locally.

Let $\Omega \subseteq \mathbb{R}^n$ be a domain with conductivity γ. If Γ is a nonempty open portion of $\partial\Omega$, the subspace of $H^{\frac{1}{2}}(\partial\Omega)$ is introduced

$$H_{co}^{\frac{1}{2}}(\Gamma) = \left\{ f \in H^{\frac{1}{2}}(\partial\Omega) \mid \operatorname{supp} f \subset \Gamma \right\}. \tag{87}$$

Definition 2. The local Dirichlet-to-Neumann map associated with γ and Γ is the operator

$$\Lambda_\gamma^\Gamma : H_{co}^{\frac{1}{2}}(\Gamma) \longrightarrow (H_{co}^{\frac{1}{2}}(\Gamma))^* \tag{88}$$

defined by

$$\langle \Lambda_\gamma^\Gamma f, \eta \rangle = \int_\Omega \gamma \nabla u \cdot \nabla \phi \, dx, \tag{89}$$

for any $f, \eta \in H_{co}^{\frac{1}{2}}(\Gamma)$, where $u \in H^1(\Omega)$ is the weak solution to

$$\begin{cases} \nabla \cdot (\gamma(x)\nabla u(x)) = 0, & \text{in} \quad \Omega, \\ u = f, & \text{on} \quad \partial\Omega, \end{cases}$$

and $\phi \in H^1(\Omega)$ is any function such that $\phi|_{\partial\Omega} = \eta$ in the trace sense.

Note that, by (89), it is easily verified that Λ_σ^Γ is self adjoint. The inverse problem is to recover γ from Λ_γ^Γ.

The procedure of reconstructing the conductivity by local measurements has been studied first by Brown [22], where the author gives a formula for reconstructing the isotropic conductivity pointwise at the boundary of a Lipschitz domain Ω without any a priori smoothness assumption of the conductivity. Nakamura and Tanuma [91] give a formula for the pointwise reconstruction of a conductivity continuous at one point x^0 of the boundary from the local D-N map when the boundary is C^1 near x^0. Under some additional regularity hypothesis the authors

give a reconstruction formula for the normal derivatives of γ on $\partial\Omega$ at $x^0 \in \partial\Omega$ up to a certain order. A direct method for reconstructing the normal derivative of the conductivity from the local Dirichlet-to-Neumann (D-N) map is presented in [92]. The result in [91] has been improved by Kang and Yun [70] to an inductive reconstruction method by using only the value of γ at x^0. The authors derive here also Hölder stability estimates for the inverse problem to identify Riemannian metrics (up to isometry) on the boundary via the local D-N map. An overview on reconstructing formulas of the conductivity and its normal derivative can be found in [93].

For related uniqueness results in the case of local boundary data, refer to Alessandrini and Gaburro [7], Bukhgeim and Uhlmann [25], Kenig et al. [73] and Isakov [68], and, for stability, [7] and Heck and Wang [55]. Recent results are also provided by Kenig and Salo [71, 72]. It is worth noting that [72] generalizes the results obtained in both [73] and [68] by making use of improved Carleman estimates with boundary terms, CGO solutions involving reflected Gaussian beam quasimodes and invertibility of (broken) geodesics ray transforms. Results of stability for cases of piecewise constant conductivities and local boundary maps have also been obtained by Alessandrini and Vessella [8], and by Di Cristo [35]. Another useful reference is [110, Sect. 7].

The Neumann-to-Dirichlet Map

In many applications of EIT especially in medical imaging, rather than the local Dirichlet-to-Neumann map, one should consider the so-called local Neumann-to-Dirichlet (N-D) map. That is, the map associating the specified current densities supported on a portion of $\Gamma \subset \partial\Omega$ to the corresponding boundary voltages, also measured on the same portion Γ of $\partial\Omega$. Usually electrodes are only applied to part of the body. Geophysics, of course, gives an extreme example where Γ is a small portion of the surface of the earth Ω. It seems appropriate at this stage to recall the definition of the N-D map and its local version for the sake of completeness [7].

Let us introduce the following function spaces (see [7])

$$H_\diamond^{\frac{1}{2}}(\partial\Omega) = \left\{ \phi \in H^{\frac{1}{2}}(\partial\Omega) | \int_{\partial\Omega} \phi = 0 \right\},$$

$$H_\diamond^{-\frac{1}{2}}(\partial\Omega) = \left\{ \psi \in H^{-\frac{1}{2}}(\partial\Omega) | \langle \psi, 1 \rangle = 0 \right\}.$$

Observe that, if considering the (global) D-N map Λ_γ, that is the map introduced in (88) Λ_γ^Γ in the special case when $\Gamma = \partial\Omega$, then it maps onto ${}_0H^{-\frac{1}{2}}(\partial\Omega)$, and, when restricted to $H_\diamond^{\frac{1}{2}}(\partial\Omega)$, it is injective with bounded inverse. Then one can define the global Neumann-to-Dirichlet map as follows.

Definition 3. The Neumann-to-Dirichlet map associated with γ, $N_\gamma : H_\diamond^{-\frac{1}{2}}(\partial\Omega) \longrightarrow H_\diamond^{\frac{1}{2}}(\partial\Omega)$ is given by

$$N_\gamma = \left(\Lambda_\gamma|_{{}_0H^{\frac{1}{2}}(\partial\Omega)}\right)^{-1}. \tag{90}$$

Note that N_γ can also be characterized as the self-adjoint operator satisfying

$$\langle \psi,\ N_\gamma \psi \rangle\ =\ \int_\Omega \gamma(x)\nabla u(x)\cdot\nabla u(x)\,dx, \tag{91}$$

for every $\psi\ \in\ H_\diamond^{-\frac{1}{2}}(\partial\Omega)$, where $u \in H^1(\Omega)$ is the weak solution to the Neumann problem

$$\begin{cases} L_\gamma u = 0, & \text{in} \quad \Omega, \\ \gamma\nabla u\cdot\nu|_{\partial\Omega} = \psi, & \text{on} \quad \partial\Omega, \\ \int_{\partial\Omega} u\ = 0. \end{cases} \tag{92}$$

At this point, it is useful to introduce the local version of the N-D map. Let Γ be an open portion of $\partial\Omega$ and let $\Delta = \partial\Omega \setminus \bar{\Gamma}$. Here, $H_{00}^{\frac{1}{2}}(\Delta)$ denotes the closure in $H^{\frac{1}{2}}(\partial\Omega)$ of the space $H_{co}^{\frac{1}{2}}(\Delta)$ previously defined in (87) and introducing

$$H_\diamond^{-\frac{1}{2}}(\Gamma) = \left\{ \psi\ \in\ H_\diamond^{-\frac{1}{2}}(\partial\Omega) |\ \langle \psi,\ f\rangle = 0, \quad \text{for any } f\ \in H_{00}^{\frac{1}{2}}(\Delta) \right\}, \tag{93}$$

that is the space of distributions $\psi \in H^{-\frac{1}{2}}(\partial\Omega)$ which are supported in $\bar{\Gamma}$ and have zero average on $\partial\Omega$. The local N-D map is then defined as follows.

Definition 4. The local Neumann-to-Dirichlet map associated with γ, Γ is the operator $N_\gamma^\Gamma :\ {}_0H^{-\frac{1}{2}}(\Gamma) \longrightarrow \left({}_0H^{-\frac{1}{2}}(\Gamma)\right)^\star \subset {}_0H^{\frac{1}{2}}(\partial\Omega)$ given by

$$\langle N_\gamma^\Gamma\ i,\ j\rangle = \langle N_\gamma\ i,\ j\rangle, \tag{94}$$

for every $i, j\ \in\ {}_0H^{-\frac{1}{2}}(\Gamma)$.

3 The Reconstruction Problem

Locating Objects and Boundaries

The simplest form of the inverse problem is to locate a single object with a conductivity contrast in a homogeneous medium. Some real situations approximate this, such as a weakly electric fish locating a single prey or the location of an insulating land mine in homogeneous soil. Typically the first test done on an EIT system experimentally is to locate a cylindrical or spherical object in a cylindrical tank. Linearization about $\gamma = 1$ simplifies to

$$\nabla^2 w = -\nabla\delta \cdot \nabla u + O\left(||\delta||^2_{L^\infty}\right). \tag{95}$$

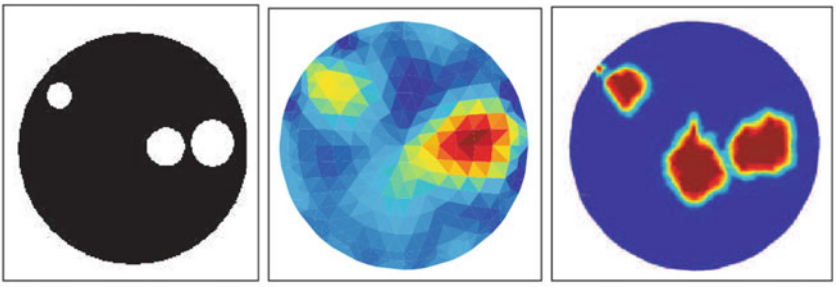
Here, the disturbance in the potential w is, to first order, the solution of Poisson's equation with a dipole source centered on the object oriented in the direction of the unperturbed electric field. With practice experienced experimenters (like electric fish) can roughly locate the object from looking at a display of the voltage changes. When it is known a priori that there is a single object, either small or with a known shape, the reconstruction problem is simply fitting a small number of model parameters (e.g., position, diameter, conductivity contrast) to the measured voltage data, using an analytical or numerical forward solver. This can be achieved using standard nonlinear optimization methods. For the two-dimensional case a fast mathematically rigorous method of locating an object (not required to be circular) from one set of Cauchy data is presented by Hanke [50]. Results on the recovery of the support of the difference between two piecewise analytic conductivities by solving the linearized problem are given in [53].

In the limiting case where the object is insulating, object location becomes a free boundary problem, where the Dirichlet-to-Neumann map is known on the known boundary and only zero Neumann data known on the unknown boundary. This is treated theoretically for example by [61] and numerically in [62]. A practical example is the location of the air core of a hydrocyclone, a device used in chemical engineering [114].

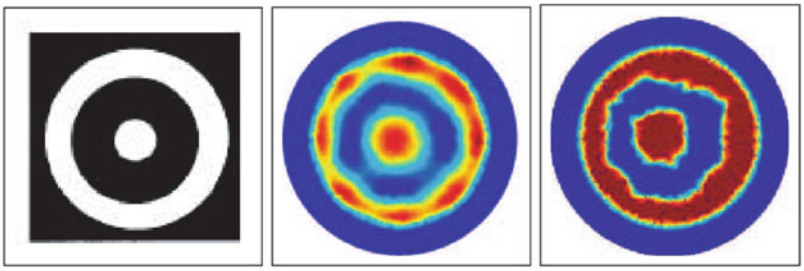
In more complex cases the conductivity may be piecewise constant with a jump discontinuity on a smooth surface. In that case there are several methods that have been tested at least on laboratory data for locating the surface of the discontinuity. One would expect in general that the location of a surface can be achieved more accurately or with less data than the recovery of a spatially varying function and this is confirmed by numerical studies.

A natural method of representing the surface of discontinuity is as a level set of a smooth surface (see chapter ▸ Level Set Methods for Structural Inversion and Image Reconstruction). This approach has the advantage that no change in parameterization is required as the number of connected components changes, in contrast for example to representing a number of star-shaped objects using spherical polar coordinates. The approach to using the level set method in EIT is exactly the same as its use in scattering problems apart from the forward problem (see chapter ▸ Inverse Scattering). Level set methods have been tested on experimental ERT and ECT data by Soleimani et al. [102]. Here some of their results are reproduced in Fig. 5.

Another approach to locating a discontinuity, common with other inverse boundary value problems for PDEs are "sampling and probe methods" in which a test is performed at each point in a grid to determine if that point is in the object. Linear sampling and factorization methods are treated in the chapter ▸ Sampling Methods. Theory and numerical results for the application of Linear Sampling to ERT for a half space are given by Hanke and Schappel [51]. Sampling methods generally require the complete transfer impedance matrix and where only incomplete measurements are available they must be interpolated.



(a) Level set reconstruction (c) from experimental ERT data for high contrast objects (a) compared with generalized Tikhonov regularization.



(b) Level set reconstruction (c) from experimental ECT data for a pipe (a) compared with generalized Tikhonov regularization.

Fig. 5 Comparison of level set reconstruction of 2D experimental data compared to generalized Tikhonov regularization using a Laplacian smoothing matrix (EIDORS-2D [113]). Thanks to Manuchehr Soleimani for reconstruction results, and Wu Quaing Yang and colleagues for the ECT data [116], the experimental ERT data was from [113]

Also in the spirit of probe methods is the *monotonicity method* of Tamburrino and Rubinacci [109]. This method follows from the observation that for γ real the map $\gamma \mapsto \mathbf{Z}_\gamma$ is monotone in the sense that $\gamma_1 \leq \gamma_2 \Rightarrow \mathbf{Z}_{\gamma_1} - \mathbf{Z}_{\gamma_2} \geq 0$, where a matrix $Z \geq 0$ if its eigenvalues are non-negative. Suppose that for some partition $\{\Omega_i\}$ of Ω (e.g., pixels or voxels)

$$\gamma = \sum_i \gamma_i \chi_{\Omega_i} \tag{96}$$

and each $\gamma_i \in \{m, M\}, 0 < m < M$. For each i let $\mathbf{Z}_i^m$ be the transfer impedance for a conductivity that is M on Ω_i and m elsewhere. Supposing $\mathbf{Z} - \mathbf{Z}_i^m$ has a negative eigenvalue, then $\gamma_i = m$. For each set in the partition the test is repeated, and it is inferred that some of the conductivity values are definitely m. The equivalent procedure is repeated for each $\mathbf{Z}_i^M$. In practice, for large enough sets in the partition and $M - m$ big enough, most conductivity values in the binary image are determined,

although this is not guaranteed. In practice the method is very fast as $\mathbf{Z}_i^m$ and $\mathbf{Z}_i^M$ can be precomputed and one only needs to find the smallest eigenvalue of two modestly sized matrices for each set in the partition. In the presence of noise, of course, one needs a sufficiently negative eigenvalue to be sure of the result of the test, and the method does assume that the conductivity is of the given form (96). Recently a partial converse of the monotonicity result has been found [54] and this promises more accurate fast methods of reconstructing the shape of an inclusion.

If conductivities on some sets are undetermined, they can then be found using other methods. For example [115] use a Markov Chain Monte Carlo method to determine the expected value and variance of undetermined pixels in ECT.

Forward Solution

Most reconstruction algorithms for EIT necessitate solution of the forward problem, that is to predict the boundary data given the conductivity. In addition, methods that use linearization typically require electric fields in the interior. The simplest case is an algorithm that uses a linear approximation calculated at a homogenous background conductivity. For simple geometries this might be done using an analytical method, while for arbitrary boundaries Boundary Element Method is a good choice, and is also suitable for the case where the conductivity is piecewise constant with discontinuities on smooth surfaces. For general conductivities the choice is between finite difference, finite volume and finite element methods (FEMs). All have been used in EIT problems. FEM has the advantage that the mesh can be adapted to a general boundary surface and to the shape and location of electrodes, whereas regular grids in finite difference/volume methods can result in more efficient computation, traded off against the fine discretization needed to represent irregular boundaries. One could also use a hybrid method such as finite element on a bounded domain of variable conductivity coupled to BEM for a homogeneous (possibly unbounded) domain.

In reconstruction methods that iteratively adjust the conductivity and resolve the forward problem, a fast forward solution is needed, whereas in methods using a linear approximation, the forward solution can be solved off-line and speed is much less important.

The simplest, and currently in EIT the most widely used, FE method is first order tetrahedral elements. Here a polyhedral approximation Ω_h to Ω is partitioned into a finite set of tetrahedra T_k, $k = 1, \ldots, n_t$ which overlap at most in a shared face, and with vertices $x_i, i = 1 < n_v$. The potential is approximated as a sum

$$u_h(x) = \sum u_i \phi_i(x), \tag{97}$$

where the ϕ_i are piecewise linear continuous functions with $\phi_i(x_j) = \delta_{ij}$. The finite element system matrix $K \in \mathbb{C}^{n_v \times n_v}$ is given by

$$K_{ij} = \int_{\Omega_p} \gamma \nabla \phi_i \cdot \nabla \phi_j \, dx. \tag{98}$$

On each tetrahedron $\nabla \phi_i$ is constant which reduces calculation of (98) in the isotropic case to the mean of γ on each tetrahedron. One then chooses an approximation to the conductivity in some space spanned by basis functions $\psi_i(x)$

$$\gamma = \sum_i \gamma_i \psi_i. \tag{99}$$

One can choose these functions to implement some a priori constraints such as smoothness and to reduce the number of unknowns in the discrete inverse problem. Or one can choose basis functions just as the characteristic functions of the tetrahedra, which makes the calculation, and updating, of the system matrix very simple. In this case, all a priori information must be incorporated later, such as by a regularization term. In general the integrals

$$\int_{\Omega_p} \psi_l \nabla \phi_i \cdot \nabla \phi_j \, dx \tag{100}$$

are evaluated using quadrature if they cannot be done explicitly. If the inverse solution uses repeated forward solutions with updated conductivity but with a fixed mesh, the coefficients (100) can be calculated once for each mesh and stored. For a boundary current density $j = \gamma \nabla u \cdot \nu$ the current vector $\mathbf{Q} \in \mathbb{R}^{n_v}$ is defined by

$$q_i = \int_{\partial\Omega} j \phi_i \, dx \tag{101}$$

and the FE system is

$$\mathbf{Ku} = \mathbf{Q}, \tag{102}$$

where $\mathbf{u}$ is the vector of u_i. One additional condition is required for a unique solution, as the voltage is only determined up to an additive constant. One way to do this is to choose one ("grounded") vertex i_g and enforce $u_{i_g} = 0$ by deleting the i_g row and column from the system (102). It is clear from (98) that for a pair of vertices indexed by i, j that are not both in any tetrahedron, $K_{ij} = 0$. The system (102) is equivalent to Ohm's and Kirchoff's law for a resistor network with resistors connecting nodes i and j when the corresponding vertices in the FE mesh share and edge (where some dihedral angles are obtuse, there is the possibility of negative conductances). It is worth noting that whatever basis is user to represent the approximate conductivity (including an anisotropic conductivity) the finite element system has only one degree of freedom per edge and one cannot hope, even with

perfect data and arithmetic, to recover more than n_e (the number of edges) unknowns from the discretization of the inverse problem.

The above formulation implements the shunt model. The CEM with specified currents can be implemented following Vauhkonen [111] using an augmented matrix. Defining

$$K^{\circ}_{ij} = K_{ij} + \sum_{l=1}^{L} \frac{1}{z_l} \int_{E_l} \phi_i \phi_j \, dx, \tag{103}$$

where, here, $|E_l|$ denotes the area of the lth electrode, and

$$K^{\partial}_{\ell\ell} = \frac{1}{z_\ell} |E_\ell| \quad \text{for} \quad \ell = 1, \ldots, L, \tag{104}$$

$$K^{\circ\partial}_{i\ell} = -\int_{E_l l} \frac{1}{z_\ell} \phi_i \, dx \qquad i = 1, .., n, \ell = 1, .., L. \tag{105}$$

The system matrix for the CEM, $\mathbf{K}^{\mathrm{CEM}} \in \mathbb{C}^{(n_v+L)\times(n_v+L)}$ is

$$\mathbf{K}^{\mathrm{CEM}} = \begin{bmatrix} K^{\circ} & K^{\circ\partial} \\ K^{\circ\partial\, T} & K^{\partial} \end{bmatrix}. \tag{106}$$

In this notation, the linear system of equations has the form

$$K^{\mathrm{CEM}} \tilde{\mathbf{u}} = \tilde{\mathbf{Q}}, \tag{107}$$

where $\tilde{\mathbf{u}} = (u_1, \ldots, u_{n_v}, V_1, .. V_L)^T$ and $\tilde{\mathbf{Q}} = (0, \ldots, 0, I_1, \ldots, I_L)^T$. The constraint $\mathbf{V} \in S$ (see section "Measurements with Electrodes") is often used to ensure uniqueness of solution. The transfer impedance matrix is obtained directly as

$$\mathbf{Z} = \left(K^{\partial} - K^{\circ\partial T} K^{\circ\,\dagger} K^{\circ\partial} \right)^{\dagger} \tag{108}$$

although it is usual to solve the system (106) as u in the interior is used in the calculation of the linearization. This formulation should only be used for reasonably large z_ℓ, as small z_ℓ will result in the block K^{∂} dominating the matrix. For an accurate forward model it is necessary to estimate the contact impedance accurately. This is more important when measurements from current carrying electrodes are used in the reconstruction, or when the electrodes are large (even if they are "passive" $I_\ell = 0$). The CEM boundary condition is rather unusual and most commercial FE systems will not include the boundary condition easily. This is one of the reasons forward solution code for EIT is generally written specifically for the purpose, such as the EIDORS project [1]. It is possible to calculate the transadmitance matrix $\mathbf{Y} = \mathbf{Z}^{\dagger}$ more easily with standard solvers. One sets Robin boundary $u + z_\ell \gamma \partial u / \partial \nu = V_\ell$ on each E_ℓ and the zero Neumann condition (4) using $\mathbf{V}$ forming a basis for S, one then takes the integral of the current over each

electrode as the current $\mathbf{I} = \mathbf{YV}$. For a given current pattern $\mathbf{I}$ one applies the Robin conditions $\mathbf{V} = \mathbf{Y}^{\dagger}\mathbf{I}$ and the solver gives the correct u. Advantages of commercial solver are that they might contain a wide variety of element types, fast solvers, and mesh generators. Disadvantages are that they may be hard to integrate as part of a nonlinear inverse solver, and it might be harder to calculate the linearization efficiently.

In fact implementing code to assemble a system matrix is quite straightforward; much harder for EIT is the generation of three-dimensional meshes. For human bodies with irregular boundaries of inaccurately known shape this is a major problem. To apply boundary conditions accurately without overfine meshes it is also important that the electrodes are approximated by unions of faces of the elements. While the accuracy of the FEM is well understood in terms of the error in the solution u, in EIT it is necessary to require that the dependence of the boundary data on the conductivity be accurate, something that is not so well understood. In addition if the conductivities vary widely it may be necessary to remesh to obtain the required accuracy, and ideally this capability with be integrated with the inverse solver [87].

Regularized Linear Methods

Methods based on linearization are popular in medical and process versions of EIT. The reasons are twofold. Process and medical applications benefit from very rapid data acquisition times with even early systems capable measuring a transfer impedance matrix in less that 0.04 s, and it was often required to produce an image in real time. The application of a precomputed (regularized) inverse of the linearized forward problem required only about $\frac{1}{2}L^2(L-1)^2$ floating point operations. For reasons of both speed and economy early systems also assumed a two-dimensional object with a single ring of electrodes arranged in a plane. The second reason for using a linear approximation is that in medical applications especially there is uncertainty in the body shape, electrode position and contact impedance. This means that a computed forward solution, based on an assumed conductivity (typically constant), has a much larger error than the errors inherent in the measurements. A compromise called *difference imaging* (by contrast to *absolute imaging*) uses a forward solution to calculate the linearization (25) and then forms an image of the difference of the conductivities between two different times, for example inspiration and expiration in a study of the lungs. Alternatively measurements can be taken simultaneously at two frequencies and a difference image formed of the permittivity.

Given a basis of applied current patterns $\mathbf{I}_i$ and a chosen set of measurements $\mathbf{M}_i$ expressed as a set of independent vectors in S that are $1/|E_l|$ for one electrode E_l, $-1/|E_k|$ for another electrode E_k (the two electrodes between which the voltage is measured), and a set of functions ψ_i with the approximate admittivity satisfying $\tilde{\gamma} = \sum \gamma_k \psi_k$, the discretization of the Fréchet derivative is the Jacobian matrix

$$J_{(ij)k} = \frac{\partial}{\partial \gamma_k} \mathrm{M}_i^T \mathrm{Z} \mathrm{I}_j = -\int_\Omega \psi_k \nabla v_i \cdot \nabla u_j \, dx, \tag{109}$$

where $L_{\tilde{\gamma}} u_i = L_{\tilde{\gamma}} v_j = 0$ (at least approximately) with u_i satisfying the CEM with current I_i and v_j with M_j. If the finite element approximation is used to solve the forward problem it has the interesting feature that the natural approximation to the Fréchet derivative in the FE context coincides with the Fréchet derivative of them FE approximation. The indices (ij) are bracketed together as they are typically "flattened" so the matrix of measurements becomes a vector and J a matrix (rather than a tensor). Let $\tilde{\mathbf{V}}$ be the vector of all voltage measurements, $\tilde{\mathbf{V}}_{\text{calc}}$ the calculated voltages, and $\boldsymbol{\gamma}$ the vector of γ_i. Our regularized least-squares version of the linearized problem is now

$$\boldsymbol{\gamma}_{\text{reg}} = \arg\min_{\boldsymbol{\gamma}} ||\mathbf{J}\boldsymbol{\gamma} - (\tilde{\mathbf{V}} - \tilde{\mathbf{V}}_{\text{calc}})||^2 + \alpha^2 \Psi(\boldsymbol{\gamma}), \tag{110}$$

where Ψ is a penalty function and α a regularization parameter. The same formulation is used for difference imaging where $\tilde{\mathbf{V}}_{\text{calc}}$ is replaced by measured data at a different time or frequency. Typical choices for Ψ are quadratic penalties such a weighted sum of squares of the γ_i, the two norm of (a discretization) a partial differential operator $\mathbf{R}$ applied to $\boldsymbol{\gamma} - \boldsymbol{\gamma}_0$, for some assumed background $\boldsymbol{\gamma}_0$. $||\mathbf{R}(\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)||^2$. Another common choice is a weighted sum of squares, i.e., L a positive diagonal matrix. In Total Variation regularization Ψ approximates $||\nabla(\gamma - \gamma_0)||_1$, and can be used where discontinuities are expected in the conductivity. Where there is a jump discontinuity on a surface (a curve in the two-dimensional case) the total variation is the integral of the absolute value of the jump over the surface (curve). The choice of regularization parameter α, the choice of penalty function, and the solution methods are covered in the chapters ▸ Linear Inverse Problems and "Total Variation in Imaging." The singular values of J are found to decay faster than exponentially (see Fig. 9), so it is a *severely illconditioned* problem and regularization is needed even for very accurate data. There are also to some extent diminishing returns in increasing the number of electrodes without also increasing the measurement accuracy.

For a quadratic penalty function the minimization problem with $\Psi(\boldsymbol{\gamma}) = ||\mathbf{R}(\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)||^2$ the solution to (110) is given by the well-known Tikhonov inversion formula

$$\boldsymbol{\gamma}_{\text{reg}} - \boldsymbol{\gamma}_0 = \left(\mathbf{J}^*\mathbf{J} + \alpha^2 \mathbf{R}^*\mathbf{R}\right)^{-1} \mathbf{J}^* \left(\tilde{\mathbf{V}} - \tilde{\mathbf{V}}_{\text{calc}}\right). \tag{111}$$

For a total variation penalty $\Psi(\boldsymbol{\gamma}) = ||\mathbf{R}(\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)||_1$ minimization is more difficult, and standard gradient based optimization methods have difficulty with the singularity in Ψ where a component of $\mathbf{R}\boldsymbol{\gamma}$ vanishes. One way around this is to use the Primal Dual Interior Point Method; for details, see [21] and for comparison of TV and a quadratic penalty applied to a difference image of the chest, see Fig. 6.

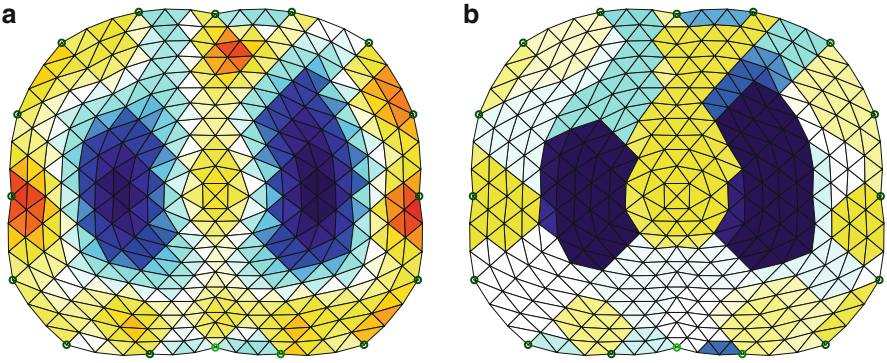


Fig. 6 Time difference EIT image of a human thorax during breathing, comparison of generalized Tikhonov $||\mathbf{R}\boldsymbol{\gamma}||_2^2$ and of the TV $||\mathbf{R}\boldsymbol{\gamma}||_1$ regularized algorithms. Both are represented on the same color scale and in arbitrary conductivity units. See [21] for details (**a**) Generalized Tikhonov (**b**) Total variation

Regularized Iterative Nonlinear Methods

As the problem is nonlinear clearly a solution based on linearization is inaccurate. Intuitively there are two aspects to the nonlinearity that are lost in a linear approximation. If one considers an object of constant conductivity away from the boundary the norm of the voltage data will exhibit a sigmoid curve as the conductivity of that object varies, as seen in the example of a concentric anomaly and illustrated numerically in Fig. 10. This means that voltage measurements *saturate*, or tends to a limiting value, as the conductivity contrast to the background tends to zero or infinity. Typically this means that linear approximations underestimate conductivity contrast. One has to be careful in communications between mathematicians and engineers: the latter will sometimes take linearity (e.g., of $\mathbf{Y}(\boldsymbol{\gamma})$) to mean a function that is homogenous of degree one, ignoring the requirement for "superposition of solutions." Considering two small spherical objects in a homogeneous background, it is known from (22) that to first order the change in u due to the objects is approximately the sum of two dipole fields. The effect of nonlinearity, the higher order terms in (22) can be thought of as interference between these two fields, analogous to higher order scattering in wave scattering problems. The practical effect is that linear approximations are poor not only at getting the correct conductivity contrast, but also at resolving a region between two objects that are close together. Many nonlinear solution methods calculate an update of the admittivity from solving a linear system, that update is applied to the conductivity in the model and the forward solution solved again. One severe problem with linear reconstruction methods that do not include a forward solver is that one cannot test if the updated admittivity even fits the data better than the initial assumption (e.g., of a constant admittivity). Such algorithms tend to produce some plausible image even if the data are erroneous.

The usual approach taken in geophysical and medical EIT to nonlinear reconstruction is to numerically perform the (nonlinear generalized Tikhonov) minimization

$$\boldsymbol{\gamma}_{\text{reg}} = \arg\min_{\boldsymbol{\gamma}} ||\tilde{\mathbf{V}}_{\text{calc}}(\boldsymbol{\gamma}) - \mathbf{V}||^2 + \alpha^2 \Psi(\boldsymbol{\gamma}) \tag{112}$$

using standard numerical optimization techniques. As the Jacobian is known explicitly it is efficient to use gradient optimization methods such as Gauss–Newton, and in that context the update step is very similar to the solution of the linear problem (110), and is a linear system for quadratic Ψ. Assuming conductivity initialized as the background level $\boldsymbol{\gamma}_0$ a typical iterative update scheme for successive approximations to the conductivity is

$$\boldsymbol{\gamma}_{n+1} = \boldsymbol{\gamma}_n + (\mathbf{J}_n^* \mathbf{J}_n + \alpha^2 Ł^* Ł)^{-1} \left(\mathbf{J}_n^* (\tilde{\mathbf{V}} - \tilde{\mathbf{V}}_{\text{calc}}(\boldsymbol{\gamma}_n) + \alpha^2 Ł^* Ł(\boldsymbol{\gamma}_0 - \boldsymbol{\gamma}_n) \right). \tag{113}$$

In contrast to the linearized problem, the nonlinear problem requires repeated solution of the forward solution $\tilde{\mathbf{V}}_{\text{calc}}(\boldsymbol{\gamma}_n)$ for variable conductivity, typically using the finite element or finite difference method. One also has to constrain the conductivity $\text{Re}\,\gamma$ to be positive and this is made easier by a choice of ϕ_i as the characteristic functions of a partition on Ω. This could be a rectangular grid or a courser tetrahedral mesh than that used for u. Accurate modeling of electrodes requires a fine discretization near electrodes, and yet one cannot hope to recover that level of detail in the admittivity near an electrode. In many practical situations a priori bounds are known for the conductivity and permittivity and as the logarithmic stability result predicts, enforcing these bounds has a stabilizing effect on the reconstruction. The positivity constraint can be enforced by a change of variables to $\log \gamma$ and this is common practice, with the Jacobian adjusted accordingly. It is generally better to perform a line search in the update direction from (113) to minimize the cost function in (112) rather than simply applying the update. Most commonly this search is approximated, for example, by fitting a few points to a low order polynomial although implementation details of this are rarely well documented. It is also worth mentioning that most absolute reconstruction algorithms start by finding a homogeneous conductivity $\boldsymbol{\gamma}_0$ best fitting the data before the iterative method starts.

In geophysical ERT nonlinear solution is well established. Although it is more common, for reasons of economy, to measure only along a line and reconstruct on the plane beneath that line, fully three-dimensional reconstruction is also widely used. The most common reconstruction code used is RES3DINV [44] which builds on the work of Loke and Barker at the University of Birmingham [84]. The code is available commercially from Loke's company Geotomo Software. RES3DINV has a finite difference forward solver used when the ground is assumed flat, and a finite element solver for known non-flat topography. In geophysical applications there is the advantage that obtaining a trangularization of the surface is a common surveying practice. The Jacobian is initialized using an analytical initial solution

assuming homogeneous conductivity. Regularized nonlinear inversion is performed using Gauss–Newton, with recalculation of Jacobian [44], or using a quasi-Newton method in which a rank one update is performed on the Jacobian. The penalty function used in the regularization is of the form $\Psi(\gamma) = ||\mathbf{R}\boldsymbol{\gamma}||_2^2$ where $\mathbf{R}$ is an approximate differential operator that penalizes horizontal and vertical variations differently. Total variation regularization $\Psi(\gamma) = ||\mathbf{R}\boldsymbol{\gamma}||_1$ is also an option in this code. When data is likely to be noisy one can select a "robust error norm," in which the one-norm is used also to measure the fit of the data to the forward solution. A maximum and minimum value of the regularization parameter can be set by the user, but in a manner similar to the classical Levenburg–Marquard method for well-posed least squares problems the parameter can be varied within that range depending on the residual at each iteration.

Although it is common in inverse problems to think of (112) as a *regularization scheme* a more rational justification for the method is probabilistic. Here, error in the measured data is considered to be sampled from a zero mean, possibly correlated, random variables. The a priori belief about the distribution of $\boldsymbol{\gamma}$ is then represented as a probability distribution. The minimization (112) is the Maximum a posteriori (MAP) estimate for the case of independent Gaussian error and with prior distribution with log probability density proportional to $-\Psi(\boldsymbol{\gamma})$. A more sophisticated approach goes beyond Gauss distributions and MAP estimates and samples the posterior distribution using Markov Chain Monte Carlo methods [69]. As this involves a large number of forward problem solutions, this is infeasible for large scale three-dimensional EIT problems. However as computers increase in speed and memory size relative to price, it is reasonable to expect this will eventually become a feasible approach. It will make it easier to approach EIT with a specific question such as "what is the volume of the region with a specified conductivity" with the answer expressed as an estimate of the probability distribution. Going back to (112) the regularization parameter α^2 controls the ratio of the variances of the prior and error distribution. In practice this choice of this parameter is somewhat subjective, and the usual techniques in choice of regularization parameter, and the caution in their application, are relevant (Fig. 7).

Results of a geophysical ERT study are shown in Fig. 8, thanks to the Geophysical Tomography Team, British Geological Survey (http://www.bgs.ac.uk/research/tomography) for this figure and the description of the survey summarized below. In this case ERT was used to identify the concentrations of *leachate*, the liquid that escapes from buried waste in a landfill site. The leachate can be extracted and recirculated to enhance the production of landfill gas, which can ultimately be used for electricity generation. It was important to use a non-invasive technique – the more standard practice of drilling exploratory wells could lead to new flow paths. Data were collected sequentially on 64 parallel survey lines, using a regular grid of electrode positions. The inter-line spacing was 15 m with a minimum electrode spacing of 5 m along line. For the current sources, electrode spacings between 15 m and 95 m were used, while electrode spacings for the measurement electrodes were between 5 m and 225 m (Figs. 9 and 10).

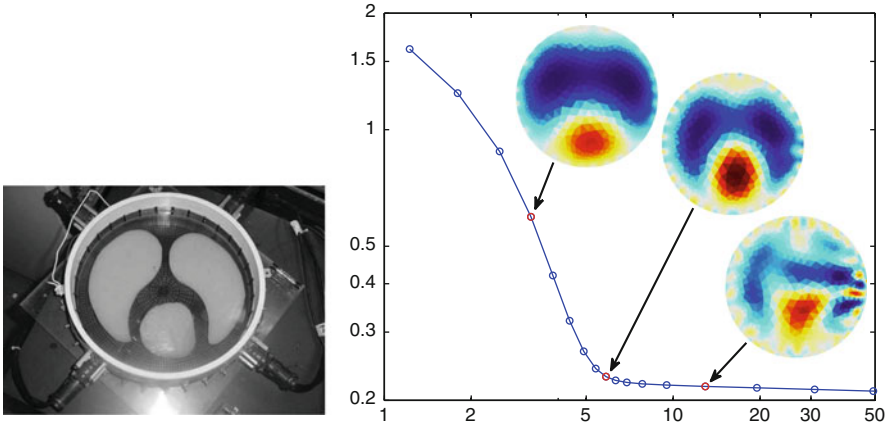


Fig. 7 An "L-curve": data mismatch $||\mathbf{V} - \mathbf{V}_{\text{calc}}(\boldsymbol{\gamma})||_2$ (*vertical*) versus regularization norm $||\mathbf{R}(\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)||_2$ (*horizontal*) for a range of six orders of magnitude of the regularization. In each case, a single step of the iterative solution was taken. Three representative images are shown illustrating the "overregularization," appropriate regularization, and "underregularization." The data are from the RPI chest phantom [65] shown *left*

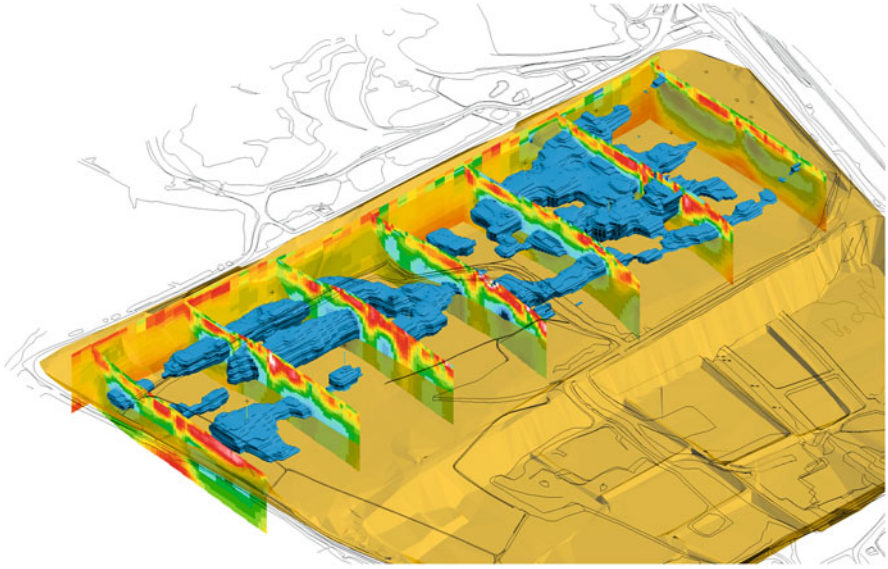


Fig. 8 A three-dimensional ERT survey of a commercial landfill site to map the volumetric distribution of leachate (*opaque blue*). Leachate is abstracted and *reproduction of any BGS materials does not amount to an endorsement by NERC or any of its employees of any product or service and no such endorsement should be stated or implied*

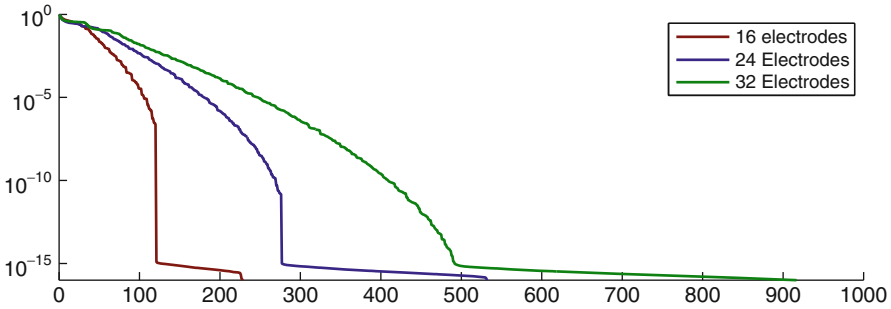


Fig. 9 Normalized singular values of the Jacobian matrix from circular 2D model with $L = 16$, 24, and 32 electrodes. EIT measurements are made with trigonometric patterns such that the number of independent measurements from L electrodes is $\frac{1}{2}(L-1)L$. Note the use of more degrees of freedom in the conductivity than the data so as to be able to study the effect of different numbers of electrodes using SVD

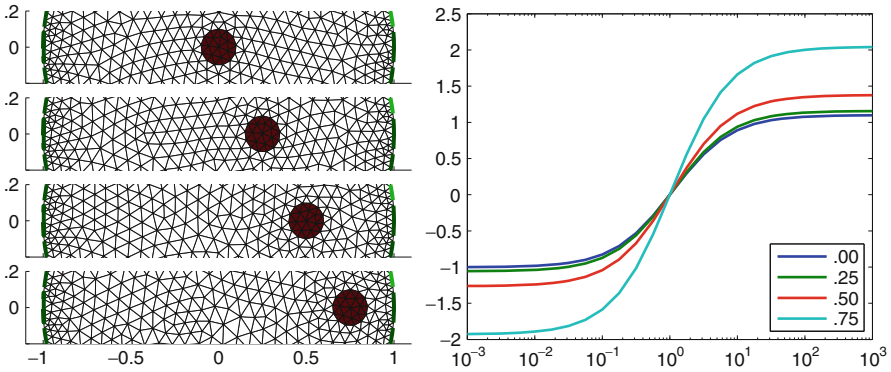


Fig. 10 Saturation of EIT signals as a function of conductivity contrast. *Left*: Slices through a finite element model of a 2D circular medium with a circular conductivity target at four horizontal positions. EIT voltages are simulated at 32 electrodes for 31 trigonometric current patterns. *Right*: change in a voltage difference as a function of conductivity contrast (target vs. background) for each *horizontal* position (*horizontal* center of contrast specified in legend). *Vertical axis* is normalized with respect to the maximum change from the central target, and scaled by the sign of conductivity change

The inversion was performed using RES3DINV with the FE forward solver with a mesh generated using the measured surface topography. The two norm was used for both penalty term and error norm. Due to the large number of datum points (approx 85,000 in total), the dataset was split in four approximately equal volumes for subsequent inversion. The resulting resistivity models were then merged to produce a final model for the entire survey area. The resulting 3D resistivity model was used to identify a total of 14 drilling locations for intrusive investigation. Eight wells were drilled and the results (i.e., initial leachate strikes within these wells) were used to calibrate the resistivity model. Based on this ground-truth

calibration a resistivity threshold value of 4 Ωm was used to represent the spatial distribution of leachate for volumetric analysis within the waste mass. A commercial visualization package was used to display cross sections, iso-resistivity surfaces as well as topography and features on the surface and the boreholes. For other similar examples of geophysical ERT see [28, 29].

Experiments on tanks in process tomography or as simulated bodies for medical EIT show that several iterations of a nonlinear method can improve the accuracy of conductivity and the shape of conductivity contours for known objects. In medical EIT it has yet to be demonstrated that the shape and electrode position can be measured and modeled with sufficient accuracy that the error in the linear approximation is greater than the modeling error. These technical difficulties are not, hopefully, insurmountable (Fig. 11).

Direct Nonlinear Solution

Nachman's [88, 89] (see also [94]) uniqueness result for the two-dimensional case was essentially constructive and has resulted in a family of reconstruction algorithms called $\bar{\partial}$-methods or scattering transform methods. Nachman's method was implemented by Siltanen et al. [100] in 2000. Of course there are few practical situations in which the two-dimensional approximation is a good one – both the conductivity and the electrodes have to be translationally invariant. Flow in a pipe with long electrodes is one example in which it is a good approximation. The main steps in the method are sketched (following Knudsen et al. [77]) and the interested reader referred to the references for details.

Assuming Ω is the unit disk for simplicity and starting with the Faddeev Green's function

$$G_k(x) := \mathrm{e}^{\mathrm{i}kx} g_k(x), \qquad g_k(x) = \frac{1}{(2\pi)^2} \int_{\mathbb{R}^2} \frac{\mathrm{e}^{\mathrm{i}x\cdot\xi}}{|\xi|^2 + 2k(\xi_1 + \mathrm{i}\xi_2)}\, d\xi \tag{114}$$

Fig. 11 (continued) Iteration (*horizontal axis*) and regularization parameter selection (*vertical axis*) for two choices of regularization matrix **R**. Data are from the RPI chest phantom [65]. The regularization parameter [α in (113)] in the *middle row* (1) was selected at the "knee" of the L-curve, indicating an appropriate level of regularization. Overregularization (*top row*) is shown for 10α, and underregularization (*bottom row*) for 0.1α. Columns indicate 1, 3, or 5 iteration of (113). With increased iteration, one can see improved separation of targets and more accurate conductivity estimates, although these improvements trade off against increased electrode artifacts due to model mismatch. The difference between the Laplacian and weighted diagonal regularization is shown in the increased smoothness of (a), especially in the underregularized case. (**a**) Laplacian regularization. (**b**) Weighted diagonal regularization

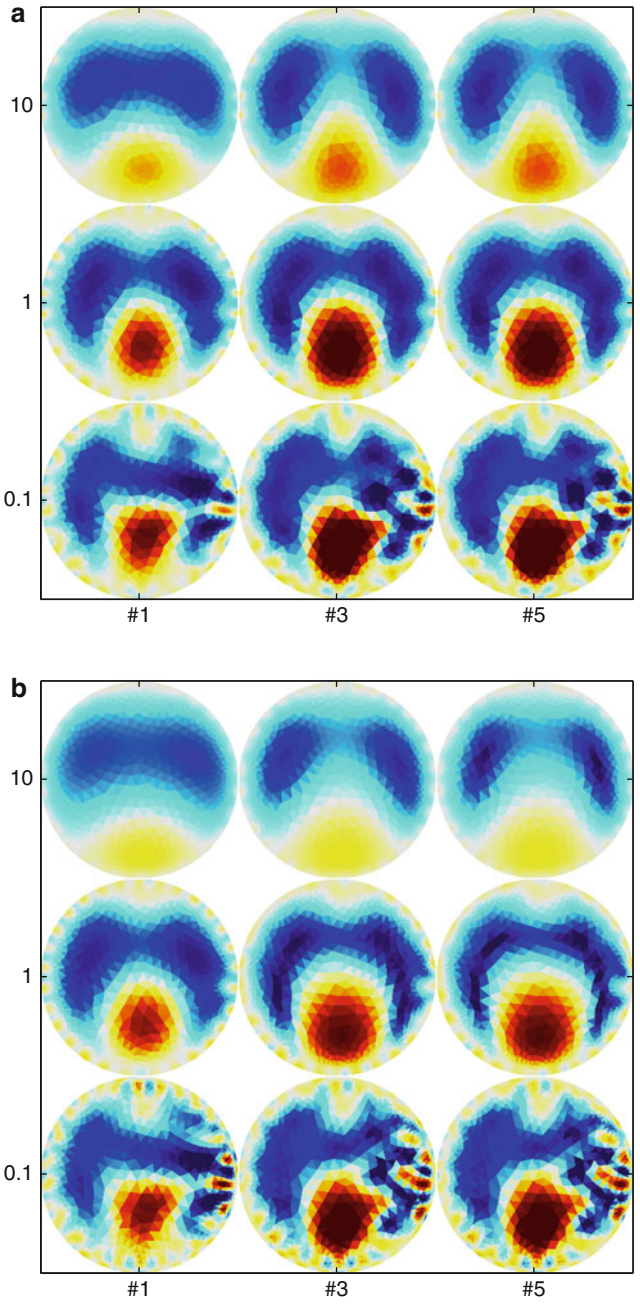


Fig. 11 (continued)

and the single layer potential

$$(S_k\phi)(x) := \int_{\partial\Omega} G_k(x-y)\phi(y)\,d\theta(y). \tag{115}$$

Here $k = k_1 + \mathrm{i}k_2$ and, by abuse of notation, considering x as a vector in $x \cdot \xi$ and a complex number $x_1 + \mathrm{i}x_2$ in the complex product kx. Here $\theta(y)$ means the angular polar coordinate of y. Here, it is assumed that the measured Dirichlet-to-Neumann map Λ_γ is available and, of course, that Λ_1 is known. The first step in the algorithm is for each fixed k to solve the linear Fredholm integral equation for a function $\psi(\cdot, k)$ on the boundary.

$$\psi(\cdot,k)|_{\partial\Omega} = \mathrm{e}^{ikx} - S_k(\Lambda_\gamma - \Lambda_1)\psi(\cdot,k)|_{\partial\Omega}. \tag{116}$$

This is an explicit calculation of the Complex Geometrics Optics solution of Theorem 3. It is fed into the calculation of what is called the *non-physical scattering transform* $\mathbf{t} : \mathbb{C} \to \mathbb{C}$ defined by

$$\mathbf{t}(k) = \int_{\partial\Omega} \mathrm{e}^{\bar{k}\bar{x}}(\Lambda_\gamma - \Lambda_1)\psi(\cdot,k)\,d\theta. \tag{117}$$

Note here that (116) is a linear equation to solve the resulting ψ depends nonlinearly on the data Λ_γ, and of course as ψ depends on the data $\mathbf{t}$ is a nonlinear function of the data. The second step is to find the conductivity from the scattering data as follows. Let $e_x(k) := \exp(\mathrm{i}(kx + \bar{k}\bar{x}))$. For each fixed x another integral equation is solved

$$V(x,k) = 1 + \frac{1}{(2\pi)^2}\int_{\mathbb{R}^2} \frac{\mathbf{t}(k')}{(k-k')\bar{k}'} e_{-x}(k')\overline{V(x,k')}\,dk_1'\,dk_2' \tag{118}$$

finally setting $\gamma(x) = V(x,0)^2$. The integral equation (118) is the solution to the partial differential equation

$$\bar{\partial}_k V(x,k) = \frac{1}{4\pi\bar{k}}\mathbf{t}(k)e_{-x}(k)\overline{V(x,k)}, \quad k \in \mathbb{C}, \tag{119}$$

where $\bar{\partial}_k = \partial/\partial\bar{k}$. Equation (119) is referred to as the $\bar{\partial}$ equation hence the name of the method.

The reconstruction procedure is therefore a direct nonlinear method in which the steps are the solution of linear equations. The only forward modeling required is the construction of Λ_1. In some practical realizations of this methods [65] an approximation to the scattering transform is used in which ψ is replaced by an exponential

$$\mathbf{t}^{\exp}(k) = \int_{\partial\Omega} e^{\bar{k}\bar{x}}(\Lambda_\gamma - \Lambda_1)\, d\theta. \tag{120}$$

In practical reconstruction schemes $\mathbf{t}$ or $\mathbf{t}^{\exp}$ are replaced by an approximation truncated to zero for $|k| > R$ for some $R > 0$, which effectively also truncates the domain of integration in (118) to the disk of radius R. Reconstruction of data from a two-dimensional agar phantom simulating a chest was performed in [65] using truncated $\mathbf{t}^{\exp}$, and in [66] a difference imaging version of the $\bar{\partial}$-method is implemented using a truncated scattering transform and applied to chest data. A rigorous regularization scheme for two-dimensional $\bar{\partial}$-reconstruction is given in [77]. In this case the regularization is applied to the *data*, in a similar spirit to X-ray CT reconstruction in which the data is filtered and then backprojected (see chapter ▸ Tomography) and the regularization is applied in the filter on the data. In this sense it is harder to understand the regularized algorithm in terms of systematic a priori information applied to the image. As in CT this is traded off against having a fast explicit reconstruction algorithm that avoids iteration.

So far the discussion of $\bar{\partial}$-methods has been confined to two-dimensional problems. At the time of writing three-dimensional direct reconstruction methods are in their infancy. A three-dimensional $\bar{\partial}$-algorithm for small conductivities is outlined in [34], and it is yet to be seen if this will result in practical implementation with noisy data on a finite array of electrodes. See the thesis of Bikowski [19] for the latest steps in this direction. If these efforts are successful, the impact on EIT is likely to be revolutionary.

4 Conclusion

Electrical impendence tomography and its relatives are among the most challenging inverse problems in imaging as the problem is nonlinear and highly illposed. The problem has inspired detailed theoretical and numerical study, and this has had an influence across a wide range of related inverse boundary value problems for (systems of) partial differential equations. Medical and industrial process applications have yet to realize their potential as routine methods while the equivalent methods in geophysics are well established. A family of direct nonlinear solution techniques until recently only valid for the two-dimensional problem may soon be extended to practical three-dimensional algorithms. If this happens fast three-dimensional nonlinear reconstruction may be possible on relatively modest computers. In some practical situations in medical and geophysical EIT the conductivity is anisotropic, in which case the solution is non-unique. A specification of the a priori information needed for a unique solution is poorly understood and practical reconstruction algorithms have yet to be proposed in the anisotropic case.

For a more complete summary of uniqueness results, the reader is referred to the review article of Uhlmann [110]. Similarly, for a review of biomedical applications of EIT, see the book by Holder [59], while subsequent progress in the medical area

can generally be found in special issues of the journal *Physiological Measurement* arising from the annual conferences on Biomedical Applications of EIT. A good reference for the details of geophysical EIT reconstruction can be found in the manual [44] and the notes by Loke [83]. For applications in process tomography see [117] and the proceedings of the biennial World Congress on Industrial Process Tomography (http://www.isipt.org/wcipt).

References

1. Adler, A., Lionheart, W.R.B.: Uses and abuses of EIDORS: an extensible software base for EIT. Physiol. Meas. **27**, S25–S42 (2006)
2. Alessandrini, G.: Stable determination of conductivity by boundary measurements. Appl. Anal. **27**, 153–172 (1988)
3. Alessandrini, G.: Singular solutions of elliptic equations and the determination of conductivity by boundary measurements. J. Differ. Equ. **84**(2), 252–272 (1990)
4. Alessandrini, G.: Determining conductivity by boundary measurements, the stability issue. In: Spigler, R. (ed.) Applied and Industrial Mathematics, pp. 317–324. Kluwer, Dordrecht (1991)
5. Alessandrini, G.: Open issues of stability for the inverse conductivity problem. J. Inverse Ill-Posed Prob. **15**, 451–460 (2007)
6. Alessandrini, G., Gaburro, R.: Determining conductivity with special anisotropy by boundary measurements. SIAM J. Math. Anal. **33**, 153–171 (2001)
7. Alessandrini, G., Gaburro, R.: The local Calderón problem and the determination at the boundary of the conductivity. Commun. PDE **34**, 918–936 (2009)
8. Alessandrini, G., Vessella, S.: Lipschitz stability for the inverse conductivity problem. Adv. Appl. Math. **35**, 207–241 (2005)
9. Ammari, H., Buffa, A., Nédélec, J.-C.: A justification of eddy currents model for the Maxwell equations. SIAM J. Appl. Math. **60**, 1805–1823 (2000)
10. Astala, K., Päivärinta, L.: Calderón's inverse conductivity problem in the plane. Ann. Math. **163**, 265–299 (2006)
11. Barber, D., Brown, B.: Recent developments in applied potential tomography – APT. In: Bacharach, S.L. (ed.) Information Processing in Medical Imaging, pp. 106–121. Nijhoff, Amsterdam (1986)
12. Barceló, J.A., Faraco, D., Ruiz, A.: Stability of the inverse problem in the plane for less regular conductivities. J. Differ. Equ. **173**, 231–270 (2001)
13. Barceló, J.A., Barceló, T., Ruiz, A.: Stability of Calderón inverse conductivity problem in the plane. J. Math. Pure. Appl. **88**, 522–556 (2007)
14. Beals, R., Coifman, R.: Transformation spectrales et equation d'evolution non lineares. Seminaire Goulaouic-Meyer-Schwarz **9**, 1981–1982 (1982)
15. Beals, R., Coifman, R.R.: Linear spectral problems, non-linear equations and the $\bar{\partial}$-method. Inverse Prob. **5**, 87–130 (1989)
16. Beretta, E., Francini, E.: Lipschitz stability for the electrical impedance tomography problem: the complex case. Commun. PDE **36**, 1723–1749 (2011)
17. Berenstein, C.A., Casadio Tarabusi, E.: Integral geometry in hyperbolic spaces and electrical impedance tomography. SIAM J. Appl. Math. **56**, 755–64 (1996)
18. Beretta, E., de Hoop, M.V., Qiu, L.: Lipschitz stability of an inverse boundary value problem for a Schrödinger-type equation. SIAM J. Math. Anal. **45**(2), 679–699 (2013)
19. Bikowski, J.: Electrical impedance tomography reconstructions in two and three dimensions; from Calderón to direct methods. Ph.D. thesis, Colorado State University, Fort Collins (2009)
20. Borcea, L.: Electrical impedance tomography. Inverse Probl. **18**, R99–R136 (2002) [Borcea, L.: Addendum to electrical impedance tomography. Inverse Prob. **19**, 997–998 (2003)]

21. Borsic, A, Graham, B.M., Adler, A., Lionheart, W.R.B.: Total variation regularization in electrical impedance tomography. IEEE Trans. Med. Imaging **29**(1), 44–54 (2010)
22. Brown, R.: Global uniqueness in the impedance-imaging problem for less regural conductivities. SIAM J. Math. Anal. **27**(4), 1049 (1996)
23. Brown, R., Torres, R.: Uniqueness in the inverse conductivity problem for conductivities with $3/2$ derivatives in L^p, $p > 2n$. J. Fourier Anal. Appl. **9**, 1049–1056 (2003)
24. Brown, R., Uhlmann, G.: Uniqueness in the inverse conductivity problem with less regular conductivities in two dimensions. Commun. PDE. **22**, 1009–1027 (1997)
25. Bukhgeim, A.L., Uhlmann, G.: Recovery a potential from partial Cauchy data. Commun. PDE. **27**, 653–668 (2002)
26. Calderón, A.P.: On an inverse boundary value problem. In: Seminar on Numerical Analysis and its Applications to Continuum Physics (Rio de Janeiro, 1980), pp. 65–73. Sociedade Brasileira de Matemática, Rio de Janeiro (1980)
27. Calderón, A.P.: On an inverse boundary value problem. Comput. Appl. Math. **25**(2–3), 133–138 (2006) [Note this reprint has some different typographical errors from the original: in particular on the first page the Dirichlet data for w is ϕ not zero]
28. Chambers, J.E., Meldrum, P.I., Ogilvy, R.D., Wilkinson, P.B.: Characterisation of a NAPL-contaminated former quarry site using electrical impedance tomography. Near Surf. Geophys. **3**, 79–90 (2005)
29. Chambers, J.E., Kuras, O., Meldrum, P.I., Ogilvy, R.D., Hollands, J.: Electrical resistivity tomography applied to geologic, hydrogeologic, and engineering investigations at a former waste-disposal site. Geophysics **71**, B231–B239 (2006)
30. Cheng, K., Isaacson, D., Newell, J.C., Gisser, D.G.:Electrode models for electric current computed tomography. IEEE Trans. Biomed. Eng. **36**, 918–24 (1989)
31. Cheney, M., Isaacson, D., Newell, J.C.: Electrical impedance tomography. SIAM Rev. **41**, 85–101 (1999)
32. Ciulli, S., Ispas, S., Pidcock, M.K.: Anomalous thresholds and edge singularities in electrical impedance tomography. J. Math. Phys. **37**, 4388 (1996)
33. Colin de VerdiÈre, Y., Gitler, I., Vertigan, D.: Réseaux électriques planaires. II. Comment. Math. Helv. **71**, 144–167 (1996)
34. Cornean, H., Knudsen, K., Siltanen, S.: Towards a D-bar reconstruction method for three dimensional EIT. J. Inverse Ill-Posed Prob. **14**, 111–134 (2006)
35. Di Cristo, M.: Stable determination of an inhomogeneous inclusion by local boundary measurements. J. Comput. Appl. Math. **198**, 414–425 (2007)
36. Dobson, D.C.: Stability and regularity of an inverse elliptic boundary value problem. Technical Report TR90-14 Rice University, Department of Mathematical Sciences (1990)
37. Doerstling, B.H.: A 3-d reconstruction algorithm for the linearized inverse boundary value problem for Maxwell's equations. Ph.D. thesis, Rensselaer Polytechnic Institute (1995)
38. Druskin, V.: The unique solution of the inverse problem of electrical surveying and electrical well-logging for piecewise-constant conductivity. Izv. Earth Phys. **18**, 51–53 (1982) (in Russian)
39. Druskin, V.: On uniqueness of the determination of the three-dimensional underground structures from surface measurements with variously positioned steady-state or monochromatic field sources. Sov. Phys.-Solid Earth **21**, 210–214 (1985) (in Russian)
40. Druskin, V.: On the uniqueness of inverse problems for incomplete boundary data. SIAM J. Appl. Math. **58**(5), 1591–1603 (1998)
41. Gaburro, R.: Sul problema inverso della tomografia da impedenza elettrica nel caso di conduttivitá anisotropa, Tesi di Laurea in Matematica, Universitá degli Studi di Trieste (1999)
42. Gaburro, R.: Anisotropic conductivity inverse boundary value problems. Ph.D. thesis, University of Manchester Institute of Science and Technology (UMIST), Manchester (2003)
43. Gaburro, R., Lionheart, W.R.B.: Recovering Riemannian metrics in monotone families from boundary data. Inverse Prob. **25**, 045004 (2009)
44. Geotomo Software: RES3DINV ver. 2.16, Rapid 3D Resistivity and IP Inversion Using the Least-Squares Method. Geotomo Software, Malaysia (2009). http://www.geoelectrical.com

45. Gisser, D.G., Isaacson, D., Newell, J.C.: Electric current computed tomography and eigenvalues. SIAM J. Appl. Math. **50**, 1623–1634 (1990)
46. Griffiths, H.: Magnetic induction tomography. Meas. Sci. Technol. **12**, 1126–1131 (2001)
47. Griffiths, H., Jossinet, J.: Bioelectric tissue spectroscopy from multi-frequency EIT. Physiol. Meas. **15**(Suppl. 2A), 29–35 (1994)
48. Haberman, B., Tataru, D.: Uniqueness in Calderón problem with Lipschitz conductivities. Duke Math. J. **162**(3), 435–625 (2013)
49. Hähner, P.: A periodic Faddeev-type solution operator. J. Differ. Equ. **128**, 300–308 (1996)
50. Hanke, M.: On real-time algorithms for the location search of discontinuous conductivities with one measurement. Inverse Prob. **24**, 045005 (2008)
51. Hanke, M., Schappel, B.: The factorization method for electrical impedance tomography in the half-space. SIAM J. Appl. Math. **68**, 907–924 (2008)
52. Hanke, M., Harrach, B., Hynöven, N.: Justification of point electrode models in electrical impedance tomography. Math. Models Methods Appl. Sci. **21**, 1395 (2011)
53. Harrach, B., Seo, JK.: Exact shape-reconstruction by one-step linearization in electrical impedance tomography. SIAM J. Math. Anal. **42**, 1505–1518 (2010)
54. Harrach, B., Ullrich, M.: Monotonicity based shape reconstruction in electrical impedance tomography. SIAM J. Math. Anal. **45**(6), 3382–3403 (2013). http://dx.doi.org/10.1137/120886984
55. Heck, H., Wang, J.-N.: Stability estimates for the inverse boundary value problem by partial Cauchy data. Inverse Prob. **22**, 1787–1796 (2006)
56. Heikkinen, L.M., Vilhunen, T., West, R.M., Vauhkonen, M.: Simultaneous reconstruction of electrode contact impedances and internal electrical properties: II. Laboratory experiments. Meas. Sci. Technol. **13**, 1855 (2002)
57. Heinrich, S., Schiffmann, H., Frerichs, A., Klockgether-Radke, A., Frerichs, I.: Body and head position effects on regional lung ventilation in infants: an electrical impedance tomography study. Intensive Care Med. **32**, 1392–1398 (2006)
58. Henderson, R.P., Webster, J.G.: An impedance camera for spatially specific measurements of the thorax. IEEE Trans. Biomed. Eng. **BME-25**(3), 250–254 (1978)
59. Holder, D.S.: Electrical Impedance Tomography Methods History and Applications. Institute of Physics Publishing, Bristol (2005)
60. Huang, S.M., Plaskowski, A., Xie, C.G., Beck, M.S.: Capacitance-based tomographic flow imaging system. Electron. Lett. **24**, 418–419 (1988)
61. Ikehata, M.: The enclosure method and its applications, Chapter 7. In: Analytic Extension Formulas and Their Applications (Fukuoka, 1999/Kyoto, 2000). International Society for Analysis, Applications and Computation, vol. 9, pp. 87–103. Kluwer Academic Publishers, Dordrecht (2001)
62. Ikehata, M., Siltanen, S.: Numerical method for finding the convex hull of an inclusion in conductivity from boundary measurements. Inverse Prob. **16**, 273–296 (2000)
63. Ingerman, D., Morrow, J.A.: On a characterization of the kernel of the Dirichlet-to-Neumann map for a planar region. SIAM J. Math. Anal. **29**, 106–115 (1998)
64. Isaacson, D.: Distinguishability of conductivities by electric current computed tomography. IEEE Trans. Med. Imaging **5**, 92–95 (1986)
65. Isaacson, D., Mueller, J.L., Newell, J., Siltanen, S.: Reconstructions of chest phantoms by the d-bar method for electrical impedance tomography. IEEE Trans. Med. Imaging **23**, 821–828 (2004)
66. Isaacson, D., Mueller, J.L., Newell, J., Siltanen, S.: Imaging cardiac activity by the D-bar method for electrical impedance tomography. Physiol. Meas. **27**, S43–S50 (2006)
67. Isakov, V.: Completeness of products of solutions and some inverse problems for PDE. J. Differ. Equ. **92**, 305–317 (1991)
68. Isakov, V.: On the uniqueness in the inverse conductivity problem with local data. Inverse Prob. Imaging **1**, 95–105 (2007)
69. Kaipio, J., Kolehmainen, V., Somersalo, E., Vauhkonen, M.: Statistical inversion and Monte Carlo sampling methods in electrical impedance tomography. Inverse Prob. **16**, 1487–1522 (2000)

70. Kang, H., Yun, K.: Boundary determination of conductivities and Riemannian metrics via local Dirichlet-to-Neumann operator. SIAM J. Math. Anal. **34**, 719–735 (2002)
71. Kenig, C., Salo, M.: Recent progress in the Calderón problem with partial data. Contemp. Math. **615**, 193–222 (2014)
72. Kenig, C., Salo, M.: The Calderón problem with partial data on manifolds and applications. Anal. PDE **6**(8), 2003–2048 (2013)
73. Kenig, C., Sjöstrand, J., Uhlmann, G.: The Calderón problem with partial data. Ann. Math. **165**, 567–591 (2007)
74. Kim, Y., Woo, H.W.: A prototype system and reconstruction algorithms for electrical impedance technique in medical body imaging. Clin. Phys. Physiol. Meas. **8**, 63–70 (1987)
75. Kohn, R., Vogelius, M.: Identification of an unknown conductivity by means of measurements at the boundary. SIAM-AMS Proc. **14**, 113–123 (1984)
76. Kohn, R., Vogelius, M.: Determining conductivity by boundary measurements II. Interior results. Commun. Pure Appl. Math. **38**, 643–667 (1985)
77. Knudsen, K., Lassas, M., Mueller, J.L., Siltanen, S.: Regularized D-bar method for the inverse conductivity problem. Inverse Prob. Imaging **3**, 599–562 (2009)
78. Lassas, M., Uhlmann, G.: Determining a Riemannian manifold from boundary measurements. Ann. Sci. École Norm. Sup. **34**, 771–787 (2001)
79. Lassas, M., Taylor, M., Uhlmann, G.: The Dirichlet-to-Neumann map for complete Riemannian manifolds with boundary. Commun. Geom. Anal. **11**, 207–222 (2003)
80. Lee, J.M., Uhlmann, G.: Determining anisotropic real-analytic conductivities by boundary measurements. Commun. Pure Appl. Math. **42**, 1097–112 (1989)
81. Lionheart, W.R.B.: Conformal uniqueness results in anisotropic electrical impedance imaging. Inverse Prob. **13**, 125–134 (1997)
82. Liu, L.: Stability estimates for the two-deimensional inverse conductivity problem. Ph.D. thesis, University of Rochester, New York (1997)
83. Loke, M.H.: Tutorial: 2-D and 3-D electrical imaging surveys. Geotomo Software (2010). http://www.geoelectrical.com
84. Loke, M.H., Barker, R.D.: Rapid least-squares inversion by a quasi-Newton method. Geophys. Prospect. **44**, 131–152 (1996)
85. Loke, M.H., Chambers, J.E., Ogilvy, R.D.: Inversion of 2D spectral induced polarization imaging data. Geophys. Prospect. **54**, 287–301 (2006)
86. Mandache, N.: Exponential instability in an inverse problem for the Schrödinger equation. Inverse Prob. **17**, 1435–1444 (2001)
87. Molinari, M., Blott, B.H., Cox, S.J., Daniell, G.J.: Optimal imaging with adaptive mesh refinement in electrical tomography. Physiol. Meas. **23**(1), 121–128 (2002)
88. Nachman, A.: Reconstructions from boundary measurements. Ann. Math. **128**, 531–576 (1988)
89. Nachman, A.: Global uniqueness for a two dimensional inverse boundary value problem. Ann. Math. **143**, 71–96 (1996)
90. Nachman, A., Sylvester, J., Uhlmann, G.: An n-dimensional Borg-Levinson theorem. Commun. Math. Phys. **115**, 593–605 (1988)
91. Nakamura, G., Tanuma, K.: Local determination of conductivity at the boundary from the Dirichlet-to-Neumann map. Inverse Prob. **17**, 405–419 (2001)
92. Nakamura, G., Tanuma, K.: Direct determination of the derivatives of conductivity at the boundary from the localized Dirichlet to Neumann map. Commun. Korean Math. Soc. **16**, 415–425 (2001)
93. Nakamura, G., Tanuma, K.: Formulas for reconstrucing conductivity and its normal derivative at the boundary from the localized Dirichlet to Neumann map. In: Hon, Y.-C., Yamamoto, M., Cheng, J., Lee, J.-Y. (eds.) The First International Conference on Inverse Problems: Recent Development in Theories and Numerics, City University of Hong Kong, Jan 2002, pp. 192–201. World Scientific, River Edge (2002)
94. Novikov, R.G.: A multidimensional inverse spectral problem for the equation $-\Delta\psi + (v(x) - Eu(x))\psi = 0$. Funktsional. Anal. i Prilozhen. **22**(4), 11–22, 96 (1988) (in Russian) [translation in Funct. Anal. Appl. 22(1988), no. 4, 263–272 (1989)]

95. Päivärinta, L., Panchenko, A., Uhlmann, G.: Complex geometrical optics solutions for Lipschitz conductivities. Rev. Mat. Iberoamericana **19**, 57–72 (2003)
96. Paulson, K., Breckon, W., Pidcock, M.: Electrode modeling in electrical-impedance tomography. SIAM J. Appl. Math. **52**, 1012–1022 (1992)
97. Polydorides, N., Lionheart, W.R.B.: A Matlab toolkit for three-dimensional electrical impedance tomography: a contribution to the electrical impedance and diffuse optical reconstruction software project. Meas. Sci. Technol. **13**, 1871–1883 (2002)
98. Seagar, A.D.: Probing with low frequency electric current. Ph.D. thesis, University of Canterbury, Christchurch (1983)
99. Seagar, A.D., Bates, R.H.T.: Full-wave computed tomography. Pt 4: low-frequency electric current CT. Inst. Electr. Eng Proc. Pt. A **132**, 455–466 (1985)
100. Siltanen, S., Mueller, J.L., Isaacson, D.: An implementation of the reconstruction algorithm of A. Nachman for the 2-D inverse conductivity problem. Inverse Prob. **16**, 681–699 (2000)
101. Soleimani, M., Lionheart, W.R.B.: Nonlinear image reconstruction for electrical capacitance tomography experimental data using. Meas. Sci. Technol. **16**(10), 1987–1996 (2005)
102. Soleimani, M., Lionheart, W.R.B., Dorn, O.: Level set reconstruction of conductivity and permittivity from boundary electrical measurements using expeimental data. Inverse Prob. Sci. Eng. **14**, 193–210 (2006)
103. Somersalo, E., Cheney, M., Isaacson, D.: Existence and uniqueness for electrode models for electric current computed tomography. SIAM J. Appl. Math. **52**, 1023–1040 (1992)
104. Soni, N.K.: Breast imaging using electrical impedance tomography. Ph.D. thesis, Dartmouth College (2005)
105. Sylvester, J.: An anisotropic inverse boundary value problem. Commun. Pure. Appl. Math. **43**, 201–232 (1990)
106. Sylvester, J., Uhlmann, G.: A uniqueness theorem for an inverse boundary value problem in electrical prospection. Commun. Pure Appl. Math. **39**, 92–112 (1986)
107. Sylvester, J., Uhlmann, G.: A global uniqueness theorem for an inverse boundary valued problem. Ann. Math. **125**, 153–169 (1987)
108. Sylvester, J., Uhlmann, G.: Inverse boundary value problems at the boundary-continuous dependence. Commun. Pure Appl. Math. **41**, 197–221 (1988)
109. Tamburrino, A., Rubinacci, G.: A new non-iterative inversion method for electrical resistance tomography. Inverse Prob. **18**, 1809–1829 (2002)
110. Uhlmann, G.: Topical review: electrical impedance tomography and Calderón's problem. Inverse Prob. **25**, 123011 (2009)
111. Vauhkonen, M.: Electrical impedance tomography and prior information. Ph.D. thesis, University of Kuopio (1997)
112. Vauhkonen, M., Karjalainen, P.A., Kaipio, J.P.: A Kalman filter approach to track fast impedance changes in electrical impedance tomography. IEEE Trans. Biomed. Eng. **45**, 486–493 (1998)
113. Vauhkonen, M., Lionheart, W.R.B., Heikkinen, L.M., Vauhkonen, P.J., Kaipio, J.P: A MATLAB package for the EIDORS project to reconstruct two-dimensional EIT images. Physiol. Meas. **22**, 107–111 (2001)
114. West, R.M., Jia, X., Williams, R.A.: Parametric modelling in industrial process tomography. Chem. Eng. J. **77**, 31–36 (2000)
115. West, R.M., Soleimani, M., Aykroyd, R.G., Lionheart, W.R.B.: Speed improvement of MCMC image reconstruction in tomography by partial linearization. Int. J. Tomogr. Stat. **4**(S06), 13–23 (2006)
116. Yang, W.Q., Spink, D.M., York, T.A., McCann, H.: An image-reconstruction algorithm based on Landweber's iteration method for electrical-capacitance tomography. Meas. Sci. Technol. **10**, 1065–1069 (1999)
117. York, T.: Status of electrical tomography in industrial applications. J. Electron. Imag. **10**, 608 (2001)

Synthetic Aperture Radar Imaging

Margaret Cheney and Brett Borden

Contents

1 Introduction 764
2 Historical Background 764
3 Mathematical Modeling 766
Scattering of Electromagnetic Waves 766
Basic Facts About the Wave Equation 766
Basic Scattering Theory 767
The Incident Field 769
Model for the Scattered Field 770
The Matched Filter 770
The Small-Scene Approximation 772
The Range Profile 773
4 Survey on Mathematical Analysis of Methods 774
Inverse Synthetic Aperture Radar (ISAR) 774
Synthetic Aperture Radar 779
Resolution for ISAR and Spotlight SAR 784
5 Numerical Methods 787
ISAR and Spotlight SAR Algorithms 787
Range Alignment 788
6 Open Problems 791
Problems Related to Unmodeled Motion 792
Problems Related to Unmodeled Scattering Physics 793
New Applications of Radar Imaging 795
7 Conclusion 796
Cross-References 796
References 796

M. Cheney (✉)
Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, NY, USA
e-mail: cheney@rpi.edu

B. Borden
Department of Physics, Naval Postgraduate School of Engineering, Monterey, CA, USA
e-mail: bhborden@nps.edu

© Springer Science+Business Media New York 2015

O. Scherzer (ed.), *Handbook of Mathematical Methods in Imaging*,
DOI 10.1007/978-1-4939-0790-8_49

Abstract

The purpose of this chapter is to explain the basics of radar imaging and to list a variety of associated open problems. After a short section on the historical background, the chapter includes a derivation of an approximate scalar model for radar data. The basics in inverse synthetic aperture radar (ISAR) are discussed, and a connection is made with the Radon transform. Two types of synthetic aperture radar (SAR), namely, spotlight SAR and stripmap SAR, are outlined. Resolution analysis is included for ISAR and spotlight SAR. Some numerical algorithms are discussed. Finally, the chapter ends with a listing of open problems and a bibliography for further reading.

1 Introduction

"Radar" is an acronym for RAdio Detection And Ranging. Radar was originally developed [7, 8, 64, 67, 72] as a technique for detecting objects and determining their positions by means of *echolocation*, and this remains the principal function of modern radar systems. However, radar systems have evolved over more than seven decades to perform an additional variety of very complex functions; one such function is imaging [9, 20–22, 26, 29, 35, 41, 59, 61].

Radar-based imaging is a technology that has been developed mainly within the engineering community. There are good reasons for this: some of the critical challenges are (1) transmitting microwave energy at high power, (2) detecting microwave energy, and (3) interpreting and extracting information from the received signals. The first two problems are concerned with the development of appropriate hardware; however, these problems have now largely been solved, although there is ongoing work to make the hardware smaller and lighter. The third problem essentially encompasses a set of mathematical challenges, and this is the area where most of the current effort is taking place.

Radar imaging shares much in common with optical imaging: both processes involve the use of electromagnetic waves to form images. The main difference between the two is that the wavelengths of radar are much longer than those of optics. Because the resolving ability of an imaging system depends on the ratio of the wavelength to the size of the aperture, radar imaging systems require an aperture many thousands of times larger than optical systems in order to achieve comparable resolution. Since kilometer-sized antennas are not practicable, fine-resolution radar imaging has come to rely on the so-called synthetic apertures in which a small antenna is used to sequentially sample a much larger measurement region.

2 Historical Background

Radar technology underwent rapid development during World War II; most of this work concerned developing methods to transmit radio waves and detect scattered

waves. The invention of synthetic aperture radar (SAR) is generally credited to Carl Wiley, of the Goodyear Aircraft Corporation, in 1951. The mid-1950s saw the development of the first operational systems, under the sponsorship of the US Department of Defense. These systems were developed by a collaboration between universities, such as the University of Illinois and the University of Michigan, together with companies such as Goodyear Aircraft, General Electric, Philco, and Varian. In the late 1960s, the National Aeronautics and Space Administration (NASA) began sponsoring unclassified work on SAR. Around this time, the first digital SAR processors were developed (earlier systems having used analogue optical processing). In 1978, the SEASAT-A satellite was sent up, and even though it operated only for 100 days, the images obtained from it were so useful that it became obvious that more such satellites were needed. In 1981, the shuttle imaging radar (SIR) series began, and many shuttle missions since then have involved radar imaging of the earth. In the 1990s, satellites were sent up by many countries (including Canada, Japan, and the European Space Agency), and SAR systems were sent to other planets and their moons, including Venus, Mars, and Titan. Since the beginning of the new millennium, more satellites have been launched, for example, the new European Space Agency satellite ENVISAT and the TerraSAR-X satellite, which was developed and launched by a (mainly European) public–private partnership.

Code letters for the radar frequency bands were originally used during wartime, and the usage has persisted. These abbreviations are listed in Table 1. The HF band usually carries radio signals; VHF carries radio and broadcast television; the UHF band carries television, navigation radar, and cell phone signals. Some radar systems operate at VHF and UHF; these are typically systems built for penetrating foliage, soil, and buildings. Most of the satellite synthetic aperture radar systems operate in the L-, S-, and C-bands. The S-band carries wireless Internet. Many military systems operate at X-band.

Table 1 Radar frequency bands

Band designation	Approximate frequency range	Approximate wavelengths
HF (high frequency)	3–30 MHz	50 m
VHF (very high frequency)	30–300 MHz	5 m
UHF (ultrahigh frequency)	300–1,000 MHz	1 m
L-band	1–2 GHz	20 cm
S-band	2–4 GHz	10 cm
C-band	4–8 GHz	5 cm
X-band	8–12 GHz	3 cm
Ku-band (under K)	12–18 GHz	2 cm
K-band	18–27 GHz	1.5 cm
Ka-band (above K)	27–40 GHz	1 cm
mm-wave	40–300 GHz	5 mm

3 Mathematical Modeling

SAR relies on a number of very specific simplifying assumptions about radar scattering phenomenology and data-collection scenarios:

1. Most imaging radar systems make use of the *start–stop approximation* [29], in which both the radar sensor and scattering object are assumed to be stationary during the time interval over which the pulse interacts with the target.
2. The target or scene is assumed to behave as a rigid body.
3. SAR imaging methods assume a linear relationship between the data and scene.

Scattering of Electromagnetic Waves

The present discussion considers only scattering from targets that are stationary.

For linear materials, Maxwell's equations can be used [34] to obtain an inhomogeneous wave equation for the electric field $\boldsymbol{\mathcal{E}}$ at time t and position $\boldsymbol{x}$:

$$\nabla^2 \boldsymbol{\mathcal{E}}(t,\boldsymbol{x}) - \frac{1}{c^2(\boldsymbol{x})} \frac{\partial^2 \boldsymbol{\mathcal{E}}(t,\boldsymbol{x})}{\partial t^2} = \boldsymbol{s}(t,\boldsymbol{x}) \tag{1}$$

and a similar equation for the magnetic field $\boldsymbol{\mathcal{B}}$. Here $c(\boldsymbol{x})$ denotes the speed of propagation of the wave (throughout the atmosphere, this speed is approximately independent of position and equal to the constant vacuum speed c) and $\boldsymbol{s}$ is a source term that, in general, can involve $\boldsymbol{\mathcal{E}}$ and $\boldsymbol{\mathcal{B}}$. For typical radar problems, the wave speed is constant in the region between the source and the scattering objects (targets) and varies only within the target volume. Consequently, here scattering objects are modeled solely via the source term $\boldsymbol{s}(t,\boldsymbol{x})$.

One Cartesian component of Eq. (1) is

$$\left(\nabla^2 - \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\right) \mathcal{E}(t,\boldsymbol{x}) = s(t,\boldsymbol{x}), \tag{2}$$

where atmospheric propagation between source and target has been assumed.

Basic Facts About the Wave Equation

A *fundamental solution* [69] of the inhomogeneous wave equation is a generalized function [30, 69] satisfying

$$\left(\nabla^2 - \frac{1}{c^2}\frac{\partial^2}{\partial_t^2}\right) g(t,\boldsymbol{x}) = -\delta(t)\delta(\boldsymbol{x}). \tag{3}$$

The solution of (3) that is useful is

$$g(t,\boldsymbol{x}) = \frac{\delta(t-|\boldsymbol{x}|/c)}{4\pi|\boldsymbol{x}|} = \int \frac{\mathrm{e}^{-\mathrm{i}\omega(t-|\boldsymbol{x}|/c)}}{8\pi^2|\boldsymbol{x}|}\,\mathrm{d}\omega, \tag{4}$$

where in the second equality the identity

$$\delta(t) = \frac{1}{2\pi}\int \mathrm{e}^{-\mathrm{i}\omega t}\,\mathrm{d}\omega \tag{5}$$

was used. The function $g(t,\boldsymbol{x})$ can be physically interpreted as the field at $(t,\boldsymbol{x})$ due to a source at the origin $\boldsymbol{x} = \boldsymbol{0}$ at time $t = 0$ and is called the *outgoing fundamental solution* or *(outgoing) Green's function*.

The Green's function [62] can be used to solve the constant-speed wave equation with *any* source term. In particular, the outgoing solution of

$$\left(\nabla^2 - \frac{1}{c^2}\frac{\partial^2}{\partial_t^2}\right) u(t,\boldsymbol{x}) = s(t,\boldsymbol{x}), \tag{6}$$

is

$$u(t,\boldsymbol{x}) = -\iint g(t-t',\boldsymbol{x}-\boldsymbol{y})s(t',\boldsymbol{y})\,\mathrm{d}t'\,\mathrm{d}\boldsymbol{y}. \tag{7}$$

In the frequency domain, the equations corresponding to (3) and (4) are

$$(\nabla^2 + k^2)G = -\delta \quad \text{and} \quad G(\omega,\boldsymbol{x}) = \frac{\mathrm{e}^{\mathrm{i}k|\boldsymbol{x}|}}{4\pi|\boldsymbol{x}|}, \tag{8}$$

where the wave number k is defined as $k = \omega/c$.

Basic Scattering Theory

In constant-wave velocity radar problems, the source $\boldsymbol{s}$ is a sum of two terms, $\boldsymbol{s} = \boldsymbol{s}^{\mathrm{in}} + \boldsymbol{s}^{\mathrm{sc}}$, where $\boldsymbol{s}^{\mathrm{in}}$ models the transmitting antenna and $\boldsymbol{s}^{\mathrm{sc}}$ models the scattering object. The solution $\boldsymbol{\mathcal{E}}$ to Eq. (1), which is written as $\boldsymbol{\mathcal{E}}^{\mathrm{tot}}$, therefore splits into two parts: $\boldsymbol{\mathcal{E}}^{\mathrm{tot}} = \boldsymbol{\mathcal{E}}^{\mathrm{in}} + \boldsymbol{\mathcal{E}}^{\mathrm{sc}}$. The first term, $\boldsymbol{\mathcal{E}}^{\mathrm{in}}$, satisfies the wave equation for the known, prescribed source $\boldsymbol{s}^{\mathrm{in}}$. This part is called the *incident* field, because it is incident upon the scatterers. The second term, $\boldsymbol{\mathcal{E}}^{\mathrm{sc}}$, is due to target scattering, and this part is called the *scattered* field. We use the same decomposition in the simplified scalar model.

One approach to finding the scattered field is to simply solve (2) directly, using, for example, numerical time-domain techniques. For many purposes, however, it is convenient to reformulate the scattering problem in terms of an integral equation.

The Lippmann–Schwinger Integral Equation

In scattering problems the source term $\boldsymbol{s}^{\mathrm{sc}}$ (typically) represents the target's *response* to an incident field. This part of the source function will generally depend on the geometric and material properties of the target and on the form and strength of the incident field. Consequently, $\boldsymbol{s}^{\mathrm{sc}}$ can be quite complicated to describe analytically, and in general, it will not have the same direction as $\boldsymbol{s}^{\mathrm{in}}$. Fortunately, for this article, it is not necessary to provide a detailed analysis of the target's response; for stationary objects consisting of linear materials, the scalar model s^{sc} is written as the time-domain convolution

$$s^{\mathrm{sc}}(t,\boldsymbol{x}) = \int v(t-t',\boldsymbol{x})\mathcal{E}^{\mathrm{tot}}(t',\boldsymbol{x})\,\mathrm{d}t', \tag{9}$$

where $v(t,\boldsymbol{x})$ is called the reflectivity function and depends on target orientation. In general, this function also accounts for polarization effects.

The expression (9) is used in (7) to express $\mathcal{E}^{\mathrm{sc}}$ in terms of the *Lippmann–Schwinger* integral equation [47]

$$\mathcal{E}^{\mathrm{sc}}(t,\boldsymbol{x}) = \int g(t-\tau,\boldsymbol{x}-z) \iint v(\tau-t',z)\mathcal{E}^{\mathrm{tot}}(t',z)\,\mathrm{d}t'\,\mathrm{d}\tau\,\mathrm{d}z. \tag{10}$$

The Lippmann–Schwinger Equation in the Frequency Domain

In the frequency domain, the electric field and reflectivity function become

$$E(\omega,\boldsymbol{x}) = \int \mathrm{e}^{\mathrm{i}\omega t}\mathcal{E}(t,\boldsymbol{x})\,\mathrm{d}t \quad \text{and} \quad V(\omega,z) = \int \mathrm{e}^{\mathrm{i}\omega t}v(t,z)\,\mathrm{d}t, \tag{11}$$

respectively. Thus the frequency-domain version of (2) is

$$\left(\nabla^2 + \frac{\omega^2}{c^2}\right) E(\omega,\boldsymbol{x}) = S(\omega,\boldsymbol{x}) \tag{12}$$

and of (10) is

$$E^{\mathrm{sc}}(\omega,\boldsymbol{x}) = -\int G(\omega,\boldsymbol{x}-z)V(\omega,z)E^{\mathrm{tot}}(\omega,z)\,\mathrm{d}z\,. \tag{13}$$

The reflectivity function $V(\omega,\boldsymbol{x})$ can display a sensitive dependence on ω [34, 36, 53]. When the target is small in comparison with the wavelength of the incident field, for example, V is proportional to ω^2 (this behavior is known as "Rayleigh scattering"). At higher frequencies (shorter wavelengths), the dependence on ω is typically less pronounced. In the so-called optical region, $V(\omega,\boldsymbol{x})$ is often approximated as being independent of ω (see, however, [56]); the optical approximation is used in this chapter, and the ω dependence is simply dropped.

In the time domain, this corresponds to $v(t,z) = \delta(t)V(z)$, and the delta function can be used to carry out the t' integration in (10).

The Born Approximation

For radar imaging, the field $\mathcal{E}^{\text{sc}}$ is measured at the radar antenna, and from these measurements, the goal is to determine V. However, both V and $\mathcal{E}^{\text{sc}}$ in the neighborhood of the target are unknown, and in (10) these unknowns are multiplied together. This nonlinearity makes it difficult to solve for V. Consequently, almost all work on radar imaging relies on the *Born* approximation, which is also known as the *weak-scattering* or *single-scattering* approximation [38, 47]. The Born approximation replaces $\mathcal{E}^{\text{tot}}$ on the right side of (10) by $\mathcal{E}^{\text{in}}$, which is known. This results in a linear formula for $\mathcal{E}^{\text{sc}}$ in terms of V:

$$\mathcal{E}^{\text{sc}}(t,\boldsymbol{x}) \approx \mathcal{E}_B(t,\boldsymbol{x}) \equiv \iint g(t-\tau,\boldsymbol{x}-z)V(z)\mathcal{E}^{\text{in}}(\tau,z)\,\mathrm{d}\tau\,\mathrm{d}z. \tag{14}$$

In the frequency domain, the Born approximation is

$$E_B^{\text{sc}}(\omega,\boldsymbol{x}) = -\int \frac{\mathrm{e}^{\mathrm{i}k|\boldsymbol{x}-z|}}{4\pi|\boldsymbol{x}-z|}V(z)E^{\text{in}}(\omega,z)\,\mathrm{d}z. \tag{15}$$

The Born approximation is very useful because it makes the imaging problem linear. It is not, however, always a good approximation; see Sect. 6.

The Incident Field

The incident field $\mathcal{E}^{\text{in}}$ is obtained by solving (2), where s^{in} is taken to be the relevant component of the current density on the source antenna and s^{sc} is zero. This article uses a simplified point-like antenna model, for which $s^{\text{in}}(t,\boldsymbol{x}) = p(t)\delta(\boldsymbol{x}-\boldsymbol{x}^0)$, where p is the waveform transmitted by the antenna. Typically p consists of a sequence of time-shifted pulses, so that $p(t) = \sum p_0(t-t_n)$.

In the frequency domain, the corresponding source for (12) is $S^{\text{in}}(\omega,\boldsymbol{x}) = P(\omega)\delta(\boldsymbol{x}-\boldsymbol{x}^0)$, where P denotes the inverse Fourier transform of p:

$$p(t) = \frac{1}{2\pi}\int \mathrm{e}^{-\mathrm{i}\omega t}P(\omega)\,\mathrm{d}\omega. \tag{16}$$

Use of (8) shows that the incident field in the frequency domain is

$$\begin{aligned} E^{in}(\omega,\boldsymbol{x}) &= -\int G(\omega,\boldsymbol{x}-\boldsymbol{y})P(\omega)\delta(\boldsymbol{y}-\boldsymbol{x}^0)\,\mathrm{d}\boldsymbol{y} \\ &= -P(\omega)\frac{\mathrm{e}^{\mathrm{i}k|\boldsymbol{x}-\boldsymbol{x}^0|}}{4\pi|\boldsymbol{x}-\boldsymbol{x}^0|}. \end{aligned} \tag{17}$$

Model for the Scattered Field

In monostatic radar systems, the transmit and receive antennas are colocated – often the same antenna is used. Use of (17) in (15) shows that the Born-approximated scattered field at the transmitter location $\boldsymbol{x}^0$ is

$$E_B^{sc}(\omega, \boldsymbol{x}^0) = P(\omega) \int \frac{\mathrm{e}^{2\mathrm{i}k|\boldsymbol{x}^0 - z|}}{(4\pi)^2 |\boldsymbol{x}^0 - z|^2} V(z)\, \mathrm{d}z. \tag{18}$$

Fourier transforming (18) results in an expression for the time-domain field:

$$\begin{aligned} \mathcal{E}_B^{sc}(t, \boldsymbol{x}^0) &= \iint \frac{\mathrm{e}^{-\mathrm{i}\omega(t - 2|\boldsymbol{x}^0 - z|/c)}}{2\pi (4\pi |\boldsymbol{x}^0 - z|)^2} P(\omega) V(z)\, \mathrm{d}\omega\, \mathrm{d}z \\ &= \int \frac{p(t - 2|\boldsymbol{x}^0 - z|/c)}{(4\pi |\boldsymbol{x}^0 - z|)^2} V(z)\, \mathrm{d}z. \end{aligned} \tag{19}$$

Under the Born approximation, the scattered field is a superposition of scattered fields from point-like targets $V(z') \propto \delta(z - z')$.

The Matched Filter

An important aspect of (19) is the $1/R^2$ geometrical decay (where $R = |\boldsymbol{x}^0 - z|$). When R is large (which it usually is), this decay factor results in a received signal that is extremely small – so small, in fact, that it can be dominated by thermal noise in the receiver. Thus it is difficult even to detect the presence of a target. Target detection is typically accomplished by means of a *matched filter* [19, 25, 50].

Below the matched filter is derived for scattering from a single fixed, point-like target. For such a target, by Eqs. (9) and (19), the signal scattered is simply a time-delayed version of the transmitted waveform:

$$s_{\mathrm{rec}}(t) = \rho s(t - \tau) + n(t),$$

where τ corresponds to the $2R/c$ delay, ρ is a proportionality factor related to the scatterer reflectivity $V(z)$ and the geometric decay $(4\pi |\boldsymbol{x}^0 - z|)^{-2}$, and n denotes noise.

The strategy is to apply a filter (convolution operator) to s_{rec} in order to improve the signal-to-noise ratio. The filter's impulse response (convolution kernel) is denoted by h, which implies that the filter output is

$$\eta(t) = (h * s_{\mathrm{rec}})(t) = \rho \eta_s(t) + \eta_n(t), \tag{20}$$

where

$$\eta_s(t) = \int h(t-t')s(t'-\tau)\,\mathrm{d}t' \quad \text{and} \quad \eta_n(t) = \int h(t-t')n(t')\,\mathrm{d}t'.$$

The signal output $\eta_s(\tau)$ at time τ should be as large as possible, relative to the noise output $\eta_n(\tau)$.

The noise is modeled as a random process. Thermal noise in the receiver is well approximated by white noise, which means that $\langle n(t)n^*(t')\rangle = N\delta(t-t')$, where N corresponds to the noise power and $\langle\cdot\rangle$ denotes expected value. Since the noise is random, so is η_n. Thus the signal-to-noise (SNR) ratio to be maximized is

$$\mathrm{SNR} = \frac{|\eta_s(\tau)|^2}{\langle|\eta_n(\tau)|^2\rangle}. \tag{21}$$

First, the denominator of (21) is

$$\begin{aligned}
\langle|\eta_n(\tau)|^2\rangle &= \left\langle \left| \int h(\tau-t')n(t')\,\mathrm{d}t' \right|^2 \right\rangle \\
&= \left\langle \int h(\tau-t')n(t')\,\mathrm{d}t' \left(\int h(\tau-t'')n(t'')\,\mathrm{d}t'' \right)^* \right\rangle \\
&= \iint h(\tau-t')h^*(\tau-t'') \underbrace{\langle n(t')n^*(t'')\rangle}_{N\delta(t'-t'')}\,\mathrm{d}t'\,\mathrm{d}t'' \\
&= N\int \left|h(\tau-t')\right|^2\,\mathrm{d}t' = N\int |h(t)|^2\,\mathrm{d}t,
\end{aligned}$$

where in the last line the change of variables $t = \tau - t'$ has been made and where the star denotes complex conjugation. Thus (21) becomes

$$\mathrm{SNR} = \frac{\left|\int h(\tau-t')s(t'-\tau)\,\mathrm{d}t'\right|^2}{N\int |h(t)|^2\,\mathrm{d}t} = \frac{\left|\int h(t)s(-t)\,\mathrm{d}t\right|^2}{N\int |h(t)|^2\,\mathrm{d}t}, \tag{22}$$

where in the numerator the change of variables $t = \tau - t'$ has been made. To the numerator of (22), the Cauchy–Schwarz inequality can be used to conclude that the numerator, and therefore the quotient (22), is maximized when h is chosen, so that

$$h(t) = s^*(-t).$$

This is the impulse response of the matched filter. Thus to obtain the best signal-to-noise ratio, the received signal should be convolved with the time-reversed, complex-conjugated version of the expected signal.

With this choice, the filter (20) can be written as

$$\eta(t) = \int h(t-t'')s_{\rm rec}(t'')\,dt'' = \int s^*(t''-t)s_{\rm rec}(t'')\,dt'' = \int s^*(t')s_{\rm rec}(t'+t)\,dt', \tag{23}$$

which is a *correlation* between s and $s_{\rm rec}$. If $s = s_{\rm rec}$, (23) is called an *autocorrelation*. Radar receivers which perform this kind of signal processing are known as "correlation receivers."

The Effect of Matched Filtering on Radar Data When applied to (18), the output of the correlation receiver is

$$\begin{aligned}
\eta(t,\boldsymbol{x}^0) &\approx \int p^*(t'-t)\mathcal{E}_B^{sc}(t',\boldsymbol{x}^0)dt' \\
&= \int\left(\frac{1}{2\pi}\int e^{i\omega'(t'-t)}P^*(\omega')d\omega'\right)\iint\frac{e^{-i\omega(t'-2|\boldsymbol{x}^0-\boldsymbol{z}|/c)}}{2\pi(4\pi|\boldsymbol{x}^0-\boldsymbol{z}|)^2}P(\omega)V(\boldsymbol{z})d\omega d\boldsymbol{z}\,dt' \\
&= \iiint \underbrace{\frac{1}{2\pi}\int e^{i(\omega-\omega)t'}dt'}_{\delta(\omega'-\omega)}\frac{e^{-i\omega(t-2|\boldsymbol{x}^0-\boldsymbol{z}|/c)}}{(4\pi|\boldsymbol{x}^0-\boldsymbol{z}|)^2}P(\omega)P^*(\omega')V(\boldsymbol{z})d\omega' d\omega d\boldsymbol{z} \\
&= \iint\frac{e^{-i\omega(t-2|\boldsymbol{x}^0-\boldsymbol{z}|/c)}}{(4\pi|\boldsymbol{x}^0-\boldsymbol{z}|)^2}|P(\omega)|^2V(\boldsymbol{z})d\omega d\boldsymbol{z}.
\end{aligned} \tag{24}$$

Thus, the effect of matched filtering is simply to replace $P(\omega)$ in the first line of (19) by $2\pi|P(\omega)|^2$.

The Small-Scene Approximation

The *small-scene* approximation, namely,

$$|\boldsymbol{x}-\boldsymbol{y}| = |\boldsymbol{x}| - \hat{\boldsymbol{x}}\cdot\boldsymbol{y} + O\left(\frac{|\boldsymbol{y}|^2}{|\boldsymbol{x}|}\right), \tag{25}$$

where $\hat{\boldsymbol{x}}$ denotes a unit vector in the direction $\boldsymbol{x}$ and is often applied to situations in which the scene to be imaged is small in comparison with its average distance from the radar. This approximation is valid for $|\boldsymbol{x}| \gg |\boldsymbol{y}|$.

Use of (25) in (4) gives rise to the large-$|\boldsymbol{x}|$ expansion of the Green's function: [12, 19]

$$G(\omega,\boldsymbol{x}-\boldsymbol{y}) = \frac{e^{ik|\boldsymbol{x}-\boldsymbol{y}|}}{4\pi|\boldsymbol{x}-\boldsymbol{y}|} = \frac{e^{ik|\boldsymbol{x}|}}{4\pi|\boldsymbol{x}|}e^{-ik\hat{\boldsymbol{x}}\cdot\boldsymbol{y}}\left(1+O\left(\frac{|\boldsymbol{y}|}{|\boldsymbol{x}|}\right)\right)\left(1+O\left(\frac{k|\boldsymbol{y}|^2}{|\boldsymbol{x}|}\right)\right). \tag{26}$$

Here, the first-order term must be included in the exponential because $k\hat{\boldsymbol{x}}\cdot\boldsymbol{y}$ can take on values that are large fractions of 2π.

Small-Scene, Matched-Filtered Radar Data In (19), the origin of coordinates can be chosen to be in or near the target, and then the small-scene expansion (26) (with z playing the role of $\boldsymbol{y}$) can be used in the matched-filtered version of (19). This results in the expression for the matched-filtered data:

$$\eta_B(t) = \frac{1}{(4\pi)^2|\boldsymbol{x}^0|^2} \iint \mathrm{e}^{-\mathrm{i}\omega(t-2|\boldsymbol{x}^0|/c+2\hat{\boldsymbol{x}}^0\cdot z/c)}|P(\omega)|^2 V(z)\,\mathrm{d}\omega\,\mathrm{d}z. \tag{27}$$

The inverse Fourier transform of (27) gives

$$D_B(\omega) = \frac{\mathrm{e}^{2\mathrm{i}k|\boldsymbol{x}^0|}}{(4\pi)^2|\boldsymbol{x}^0|^2}|P(\omega)|^2 \underbrace{\int \mathrm{e}^{-2\mathrm{i}k\hat{\boldsymbol{x}}^0\cdot z} V(z)\,\mathrm{d}z}_{\mathcal{F}[V](2k\hat{\boldsymbol{x}}^0)}. \tag{28}$$

Thus we see that each frequency component of the data provides us with a Fourier component of the reflectivity V.

The Range Profile

Signals with large bandwidth are commonly used in synthetic aperture imaging. When the bandwidth is large, the pulse p is said to be a *high-range-resolution* (HRR) pulse. An especially simple large-bandwidth signal is one for which $|P(\omega)|^2$ is constant over its support. In this case, the ω-integral in Eq. (27) reduces to a scaled $\mathrm{sinc}(t)$ function centered on

$$t = 2|\boldsymbol{x}^0|/c + 2\hat{\boldsymbol{x}}^0\cdot z/c,$$

and the width of this sinc function is inversely proportional to the bandwidth. When the support of $|P(\omega)|^2$ is infinite, of course, this $\mathrm{sinc}(t)$ becomes a delta function. Thus large-bandwidth (HRR), matched-filtered data can be approximated by

$$\eta_B(t) \approx \frac{1}{(4\pi)^2|\boldsymbol{x}^0|^2} \int \delta(t - 2|\boldsymbol{x}^0|/c + 2\hat{\boldsymbol{x}}^0\cdot z/c) V(z)\,\mathrm{d}z. \tag{29}$$

Since time delay and range are related in monostatic radar systems as $t = 2R/c$, Eq. (29) can be seen to be a relation between the radar data $\eta_B(t)$ and the integral of the target reflectivity function over the plane

$$R = |\boldsymbol{x}^0| + \hat{\boldsymbol{x}}^0\cdot z$$

(with respect to the radar). Such data are said to form a "range profile" of the target. An example of an HRR range profile is displayed in Fig. 1.

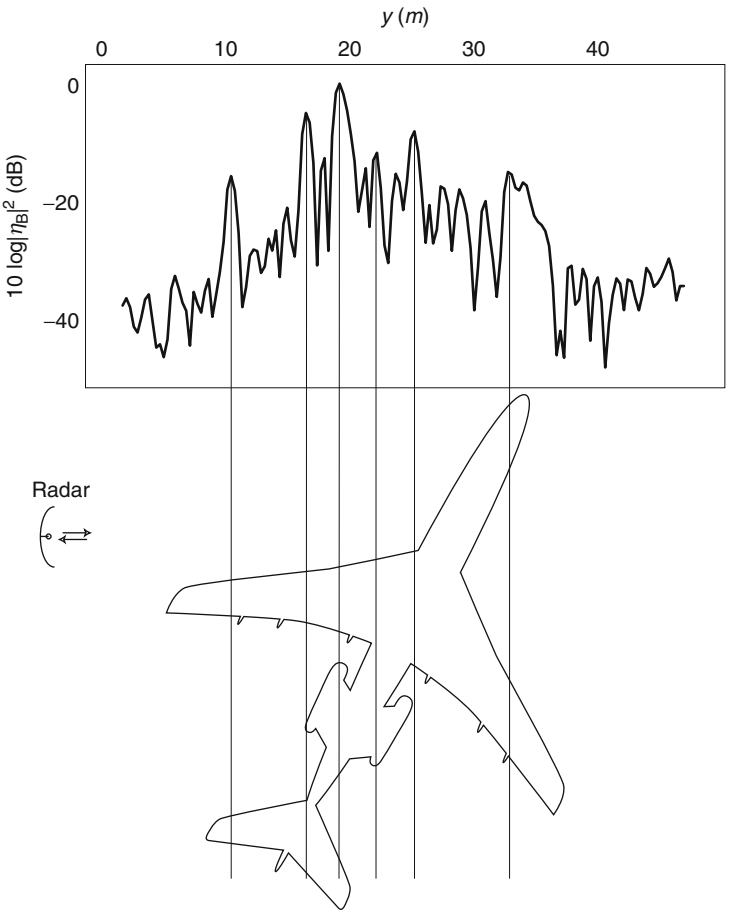


Fig. 1 Example of an HRR range profile of an aircraft (orientation displayed in *inset*)

4 Survey on Mathematical Analysis of Methods

The mathematical models discussed above assume that the target $V(\boldsymbol{z})$ is stationary during its interaction with a radar pulse. However, synthetic aperture imaging techniques assume that the target moves with respect to the radar *between* pulses.

Inverse Synthetic Aperture Radar (ISAR)

A fixed radar system staring at a rotating target is equivalent (by change of reference frame) to a stationary target viewed by a radar moving (from pulse to pulse) on a circular arc. This circular arc will define, over time, a synthetic

aperture, and sequential radar pulses can be used to sample those data that would be collected by a much larger radar antenna. Radar imaging based on such a data-collection configuration is known as *inverse synthetic aperture radar* (ISAR) imaging [5, 15, 41, 59, 66, 74]. This imaging scheme is typically used for imaging airplanes, spacecraft, and ships. In these cases, the target is relatively small and usually isolated.

Modeling Rotating Targets The target reflectivity function in a frame fixed to the target is denoted by q. Then, as seen by the radar, the reflectivity function is $V(\boldsymbol{x}) = q(\mathcal{O}(\theta_n)\boldsymbol{x})$, where $\mathcal{O}$ is an orthogonal matrix and where $t_n = \theta_n$ denotes the time at the start of the nth pulse of the sequence.

For example, if the radar is in the plane perpendicular to the axis of rotation (so-called turntable geometry), then the orthogonal matrix $\mathcal{O}$ can be written as

$$\mathcal{O}(\theta) = \begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{30}$$

and $V(\boldsymbol{x}) = q(x_1\cos\theta - x_2\sin\theta, x_1\sin\theta + x_2\cos\theta, x_3)$.

Radar Data from Rotating Targets The use of $V(\boldsymbol{x}) = q\left(\mathcal{O}(\theta_n)\boldsymbol{x}\right)$ in (28) provides a model for the data from the nth pulse:

$$D_B(\omega, \theta_n) = \frac{\mathrm{e}^{2ik|\boldsymbol{x}^0|}}{(4\pi)^2|\boldsymbol{x}^0|^2}|P_0(\omega)|^2 \int \mathrm{e}^{-2ik\hat{\boldsymbol{x}}^0\cdot z} q(\underbrace{\mathcal{O}(\theta_n)\boldsymbol{z}}_{\boldsymbol{y}})\,\mathrm{d}\boldsymbol{z}. \tag{31}$$

In (31), the change of variables $\boldsymbol{y} = \mathcal{O}(\theta_n)\boldsymbol{z}$ is made. Then use is made of the fact that the inverse of an orthogonal matrix is its transpose, which means that $\boldsymbol{x}^0 \cdot \mathcal{O}^{-1}(\theta_n)\boldsymbol{y} = \mathcal{O}(\theta_n)\boldsymbol{x}^0 \cdot \boldsymbol{y}$. The result is that (31) can be written in the form

$$D_B(\omega, \theta_n) = \frac{\mathrm{e}^{2ik|\boldsymbol{x}^0|}}{(4\pi)^2|\boldsymbol{x}^0|^2}|P_0(\omega)|^2 \underbrace{\int \mathrm{e}^{-2ik\mathcal{O}(\theta_n)\hat{\boldsymbol{x}}^0\cdot y} q(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y}}_{\propto \mathcal{F}[q](2k\mathcal{O}(\theta_n)\hat{\boldsymbol{x}}^0)}. \tag{32}$$

Thus, the frequency-domain data are proportional to the inverse Fourier transform of q, evaluated at points in a domain defined by the angles of the sampled target orientation and the radar bandwidth (see Fig. 2). Consequently, a Fourier transform produces a target image.

The target rotation angle is usually not known. However, if the target is rotating with constant angular velocity, the image produced by the Fourier transform gives rise to a stretched or contracted image, from which the target is usually recognizable [5, 41, 66, 72].

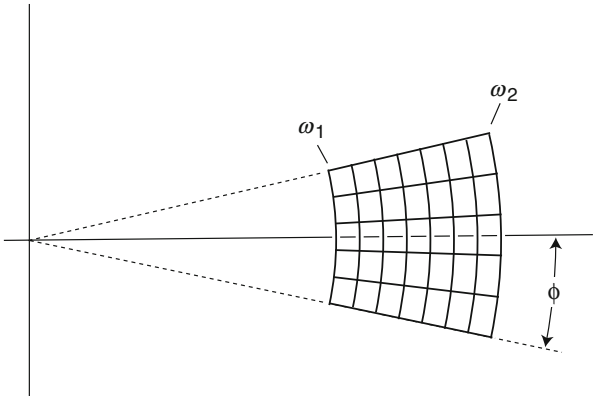


Fig. 2 The data-collection manifold for turntable geometry

The Data-Collection Manifold

The Fourier components of the target that can be measured by the radar are those in the set

$$\Omega_z = \{2k\mathcal{O}(\theta_n)\hat{\boldsymbol{x}}^0\}, \tag{33}$$

where n ranges over the indices of pulses for which the point $\boldsymbol{z}$ is in the antenna beam and where $k = \omega/c$ with ω ranging over the angular frequencies received by the radar receiver. The region determined in this manner is called the *data-collection manifold*. The extent of the set of angles is called the *synthetic aperture*, and the extent of the set of frequencies is called the *bandwidth*. Typical synthetic apertures are on the order of a few degrees, and bandwidths of $2\pi \times 500 \times 10^6$ rad/s are not uncommon. Figure 2 shows an example of data-collection manifold corresponding to turntable geometry; Fig. 3 shows others that correspond to more complex motion. Typical SAR data-collection manifolds are two-dimensional manifolds. The larger the data-collection manifold at $\boldsymbol{z}$, the better the resolution at $\boldsymbol{z}$.

Examples of ISAR images are shown in Figs. 4 and 5.

ISAR in the Time Domain

Fourier transforming (32) into the time-domain results in

$$\eta_B(t,\theta_n) \propto \iint \mathrm{e}^{-\mathrm{i}\omega(t-2|\boldsymbol{x}^0|/c+2\mathcal{O}(\theta_n)\hat{\boldsymbol{x}}^0\cdot\boldsymbol{y}/c)}|P_0(\omega)|^2\,\mathrm{d}\omega\; q(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y}. \tag{34}$$

Evaluation of η_B at a shifted time results in the simpler expression

$$\eta_B\left(t+\frac{2|\boldsymbol{x}^0|}{c},\theta_n\right) = \iint \mathrm{e}^{-\mathrm{i}\omega(t+2\mathcal{O}(\theta_n)\hat{\boldsymbol{x}}\cdot\boldsymbol{y}/c)}|P_0(\omega)|^2\,\mathrm{d}\omega\; q(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y}. \tag{35}$$

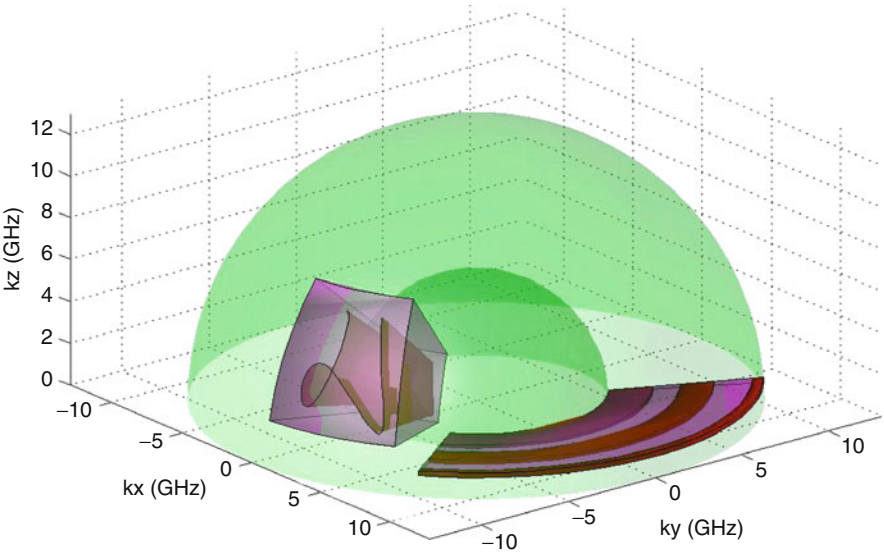


Fig. 3 The dark surfaces represent some typical data-collection manifolds that are subsets of a more complete "data dome"

With the temporary notation $\tau = -2\mathcal{O}(\theta_n)\hat{\boldsymbol{x}} \cdot \boldsymbol{y}/c$, the ω integral on the right side of (35) can be written as

$$\int \mathrm{e}^{-\mathrm{i}\omega(t-\tau)} |P_0(\omega)|^2 \, \mathrm{d}\omega = \int \delta(s-\tau)\beta(t-s) \, \mathrm{d}s, \tag{36}$$

where

$$\beta(t-s) = \int \mathrm{e}^{-\mathrm{i}\omega(t-s)} |P_0(\omega)|^2 \, \mathrm{d}\omega.$$

With (36), η_B can be written

$$\begin{aligned} \eta_B\left(t + \frac{2|\boldsymbol{x}^0|}{c}, \theta_n\right) &= \int \beta(t-s) \int \delta\left(s + \frac{2\mathcal{O}(\theta_n)\hat{\boldsymbol{x}}}{c} \cdot \boldsymbol{y}\right) q(\boldsymbol{y}) \, \mathrm{d}\boldsymbol{y} \, \mathrm{d}s \\ &= \beta * \mathcal{R}[q]\left(\frac{-2\mathcal{O}(\theta_n)\hat{\boldsymbol{x}}}{c}\right), \end{aligned}$$

where

$$\mathcal{R}[q](s, \hat{\boldsymbol{\mu}}) = \int \delta(s - \hat{\boldsymbol{\mu}} \cdot \boldsymbol{y}) q(\boldsymbol{y}) \, \mathrm{d}\boldsymbol{y} \tag{37}$$

is the *Radon transform* [43, 46]. Here $\hat{\boldsymbol{\mu}}$ denotes a unit vector. In other words, the Radon transform of q is defined as the integral of q over the plane $s = \hat{\boldsymbol{\mu}} \cdot \boldsymbol{y}$.

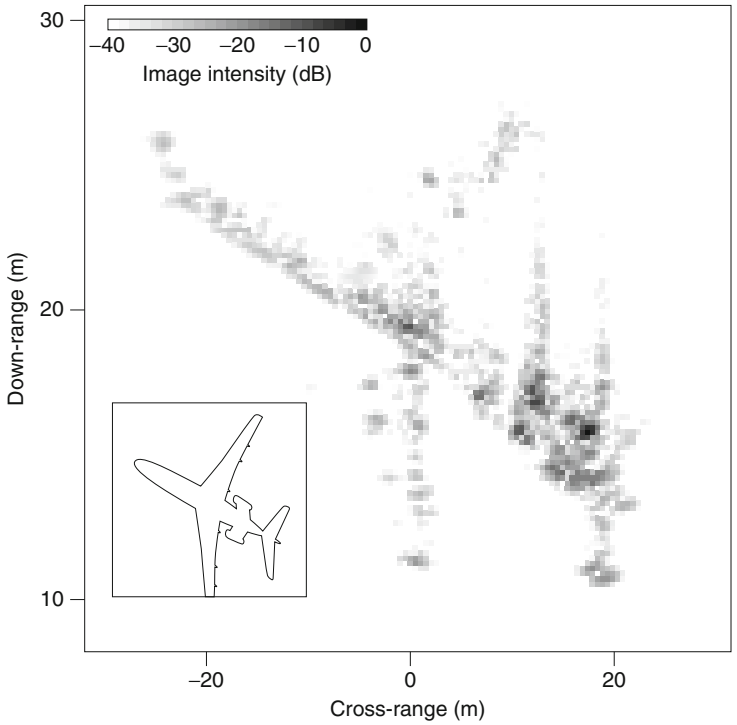


Fig. 4 An ISAR image of a Boeing 727 from a 5° aperture [70]

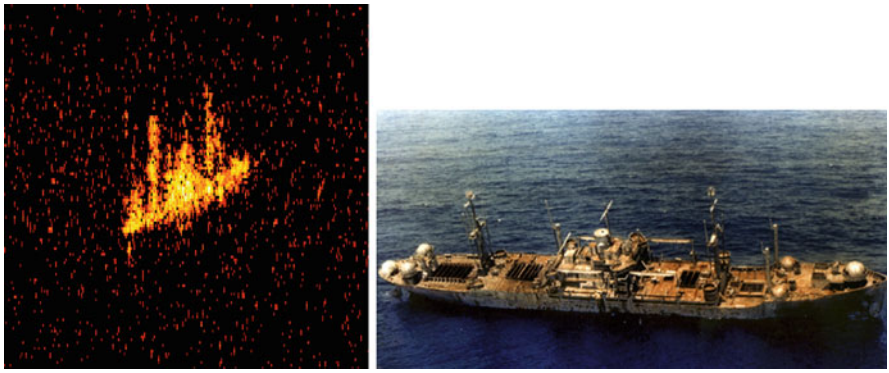


Fig. 5 On the *left* is an ISAR image of a ship; on the *right* is an optical image of the same ship (Courtesy Naval Research Laboratory)

ISAR systems typically use a high-range-resolution (large-bandwidth) waveform, so that $\beta \approx \delta$ (see section "The Range Profile"). Thus ISAR imaging from time-domain data becomes a problem of inverting the Radon transform.

Synthetic Aperture Radar

In ISAR, the target rotates and the radar is stationary, whereas in synthetic aperture radar (SAR), the target is stationary and the radar moves. (In typical ISAR data-collection scenarios, both the radar and the target are actually in motion, and so this distinction is somewhat arbitrary.) For most SAR systems [9, 20, 21, 29, 35, 60], the antenna is pointed toward the earth. For an antenna viewing the earth, an antenna beam pattern must be included in the model. For highly directive antennas, often simply the antenna "footprint," which is the illuminated area on the ground, is used.

For a receiving antenna at the same location as the transmitting antenna, the scalar Born model for the received signal is

$$S_B(\omega) = \int \mathrm{e}^{2\mathrm{i}k|\boldsymbol{x}^0-\boldsymbol{y}|} A(\omega, \boldsymbol{x}^0, \boldsymbol{y}) V(\boldsymbol{y})\, \mathrm{d}\boldsymbol{y}, \tag{38}$$

where A incorporates the geometrical spreading factors $|\boldsymbol{x}^0 - \boldsymbol{y}|^{-2}$, transmitted waveform, and antenna beam pattern. More details can be found in [15].

SAR data-collection systems are usually configured to transmit a series of pulses with the nth pulse transmitted at time t_n. The antenna position at time t_n is denoted by $\boldsymbol{\gamma}_n$. Because the time scale on which the antenna moves is much slower than the time scale on which the electromagnetic waves propagate, the time scales separate into a *slow time*, which corresponds to the n of t_n, and a *fast time* t.

In (38) the antenna position $\boldsymbol{x}^0$ is replaced by $\boldsymbol{\gamma}_n$:

$$D(\omega, n) = F[V](\omega, s) \equiv \int \mathrm{e}^{2\mathrm{i}k|\boldsymbol{\gamma}_n-\boldsymbol{y}|} A(\omega, n, \boldsymbol{y}) V(\boldsymbol{y})\, \mathrm{d}\boldsymbol{y}, \tag{39}$$

where with a slight abuse of notation, the $\boldsymbol{x}^0$ in the argument of A has been replaced by n. This notation also allows for the possibility that the waveform and antenna beam pattern could be different at different points along the flight path. The time-domain version of (39) is

$$d(t, n) = \int \mathrm{e}^{-i\omega[t-2|\boldsymbol{\gamma}_n-\boldsymbol{y}|/c]} A(\omega, n, \boldsymbol{y}) V(\boldsymbol{y})\, \mathrm{d}\boldsymbol{y}. \tag{40}$$

The goal of SAR is to determine V from the data d.

As in the case of ISAR, assuming that $\boldsymbol{\gamma}$ and A are known, the data depend on two variables, so it should be possible to form a two-dimensional image. For typical radar frequencies, most of the scattering takes place in a thin layer at the surface. The ground reflectivity function V is therefore assumed to be supported on

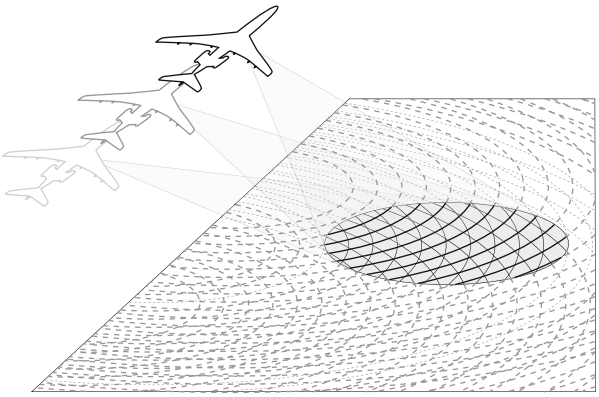


Fig. 6 In spotlight SAR, the radar is trained on a particular location as the radar moves. In this figure, the equi-range circles (*dotted lines*) are formed from the intersection of the radiated spherical wave front and the surface of a (flat) earth

a known surface. For simplicity this surface is assumed to be a flat plane, so that $V(\boldsymbol{x}) = V(\mathbf{x})\delta(x_3)$, where $\mathbf{x} = (x_1, x_2)$.

SAR imaging comes in two basic varieties: *spotlight* SAR [9, 35] and *stripmap* SAR [20, 21, 29, 60].

Spotlight SAR

Spotlight SAR is illustrated in Fig. 6. Here, the moving radar system stares at a specific location (usually on the ground), so that at each point in the flight path the same scene is illuminated from a different direction. When the ground is assumed to be a horizontal plane, the iso-range curves are large circles whose centers are directly below the antenna at $\boldsymbol{\gamma}_n$. If the radar antenna is highly directional and the antenna footprint is sufficiently far away, then the circular arcs within the footprint can be approximated as lines. Consequently, the imaging method is mathematically the same as that used in ISAR.

In particular, the origin of coordinates is taken within the footprint, and the small-scene expansion is used, which results in an expression for the matched-filtered frequency-domain data:

$$D(\omega, n) = \mathrm{e}^{2\mathrm{i}k|\boldsymbol{\gamma}_n|} \int \mathrm{e}^{2\mathrm{i}k\widehat{\boldsymbol{\gamma}}_n \cdot \boldsymbol{y}} V(\boldsymbol{y}) A(\omega, n, \boldsymbol{y})\, \mathrm{d}\boldsymbol{y}. \tag{41}$$

Within the footprint, A is approximated as a product $A = A_1(\omega, n)A_2(\boldsymbol{y})$. The function A_1 can be taken outside the integral; the function A_2 can be divided out after inverse Fourier transforming.

As in the ISAR case, the time-domain formulation of spotlight SAR leads to a problem of inverting the Radon transform. An example of a spotlight SAR image is shown in Fig. 12.

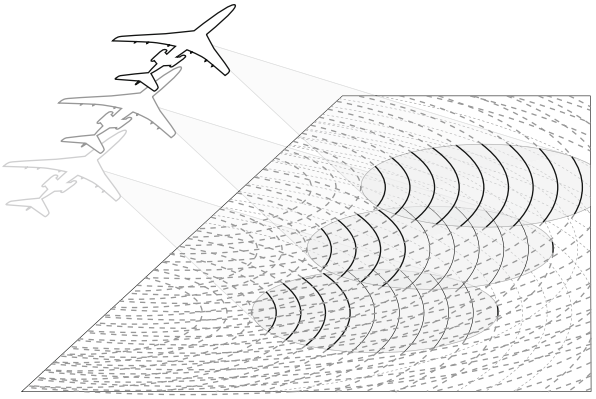


Fig. 7 Stripmap SAR acquires data without staring. The radar typically has fixed orientation with respect to the flight direction, and the data are acquired as the beam footprint sweeps over the ground

Stripmap SAR

Stripmap SAR sweeps the radar beam along with the platform without staring at a particular location on the ground (Fig. 7). The equi-range curves are still circles, but the data no longer depend only on the direction from the antenna to the scene. Moreover, because the radar does not stare at the same location, there is no natural origin of coordinates for which the small-scene expansion is valid.

To form a stripmap SAR image, the expression (39) must be inverted without the help of the small-scene approximation. One strategy is to use a filtered adjoint of the forward map F defined by Eq. (39).

The Formal Adjoint of F The adjoint $F^\dagger$ is an operator such that

$$\langle f, Fg\rangle_{\omega,s} = \langle F^\dagger f, g\rangle_{\boldsymbol{x}}, \tag{42}$$

where $\langle \cdot, \cdot \rangle$ denotes inner product in the appropriate variables. More specifically, (42) can be written as

$$\int f(\omega, s)\,(Fg)^*(\omega, s)\,\mathrm{d}\omega\,\mathrm{d}s = \int (F^\dagger f)(\boldsymbol{x}) g^*(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}. \tag{43}$$

Use of (39) in (43) and an interchange of the order of integration leads to

$$F^\dagger f(\boldsymbol{x}) = \iint \mathrm{e}^{-2\mathrm{i}k|\boldsymbol{\gamma}(s)-\boldsymbol{x}|} A(\omega, s, \boldsymbol{x})\, f(\omega, s)\,\mathrm{d}\omega\,\mathrm{d}s. \tag{44}$$

The Imaging Operator Thus, the imaging operator is assumed to be of the form

$$I(\mathbf{z}) = B[D](\mathbf{z}) \equiv \iint \mathrm{e}^{-2\mathrm{i}k|\boldsymbol{\gamma}(s)-\mathbf{z}_T|} Q(\omega, s, \mathbf{z}) D(\omega, s)\, \mathrm{d}\omega\, \mathrm{d}s, \tag{45}$$

where $\mathbf{z}_T = (\mathbf{z}, 0)$ and Q is a filter to be determined below. The time-domain version is

$$I(\mathbf{z}) = \mathcal{B}[d](\mathbf{z}) \equiv \iint \mathrm{e}^{\mathrm{i}\omega(t-2|\boldsymbol{\gamma}(s)-\mathbf{z}_T|/c)} Q(\omega, s, \mathbf{z})\, \mathrm{d}\omega d(t, s)\, \mathrm{d}s\, \mathrm{d}t. \tag{46}$$

If the filter Q were to be chosen to be identically 1, then, because of (5), the time-domain inversion would have the form

$$\begin{aligned} I(\mathbf{z}) &= \iint \delta(t - 2|\boldsymbol{\gamma}(s) - \mathbf{z}_T|/c) d(t, s)\, \mathrm{d}s\, \mathrm{d}t \\ &= \int d(2|\boldsymbol{\gamma}(s) - \mathbf{z}_T|/c, s)\, \mathrm{d}s, \end{aligned} \tag{47}$$

which can be interpreted as follows: at each antenna position s, the data is backprojected (smeared out) to all the locations $\mathbf{z}$ that are at the correct travel time $2|\boldsymbol{\gamma}(s) - \mathbf{z}_T|/c$ from the antenna location $\boldsymbol{\gamma}(s)$. Then all the contributions are summed coherently (i.e., including the phase). Figure 8 shows the partial sums over s as the antenna (white triangle) moves along a straight flight path from bottom to top.

An alternative interpretation is that to form the image at the reconstruction point $\mathbf{z}$, all the contributions from the data at all points (t, s) for which $t = 2|\boldsymbol{\gamma}(s) - \mathbf{z}_T|/c$ are coherently summed.

Note the similarity between (47) and (61): (61) backprojects over lines or planes, whereas (47) backprojects over circles. The inversion (46) first applies the filter Q and then backprojects.

Other SAR Algorithms The image formation algorithm discussed here is filtered backprojection. This algorithm has many advantages, one of them being great flexibility. This algorithm can be used for any antenna beam pattern, for any flight path, and for any waveform; a straightforward extension [48] can be used in the case when the topography is not flat.

Nevertheless, there are various other algorithms that can be used in special cases, for example, if the flight path is straight, if the antenna beam is narrow, or if a chirp waveform is used. Discussions of these algorithms can be found in the many excellent radar imaging books such as [9, 20, 29, 35, 61].

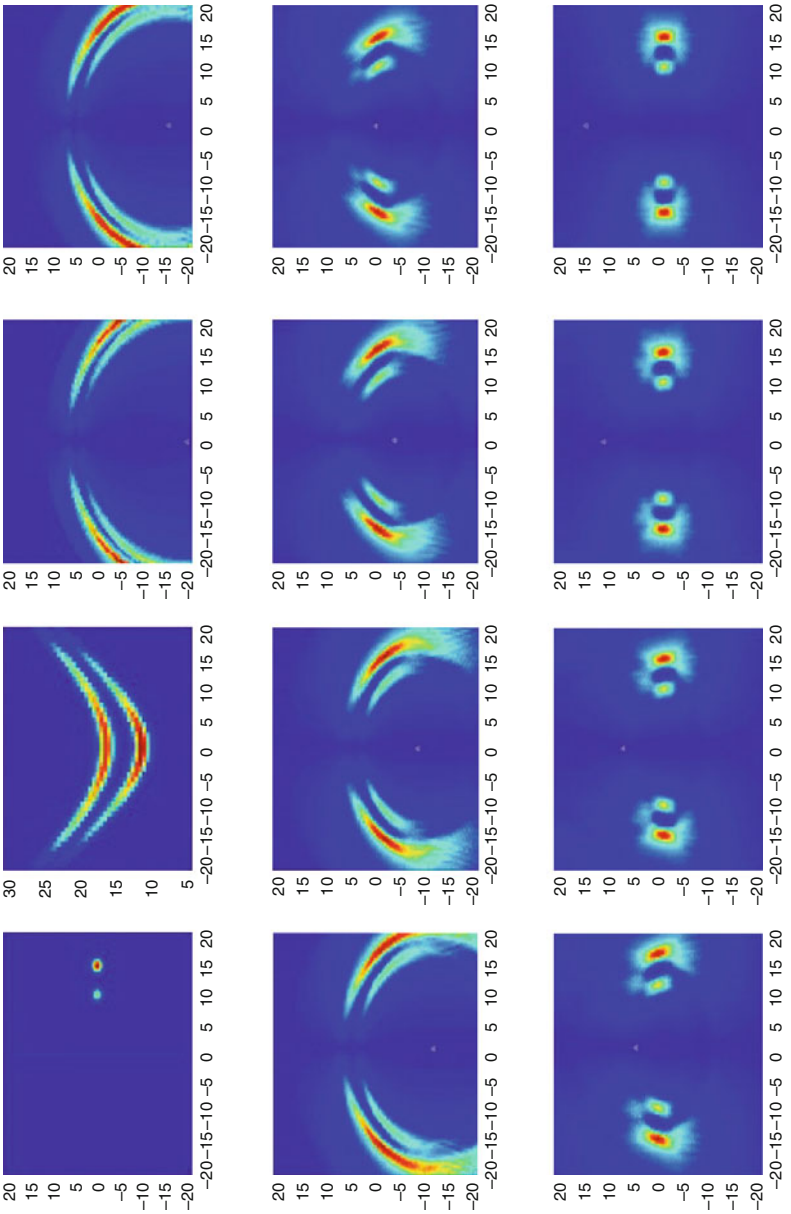


Fig. 8 This shows successive steps in the backprojection procedure for a straight flight path and an isotropic antenna. The first image is the true scene; the second is the magnitude of the data. The successive images show the image when the antenna has traveled as far as the location indicated by the small triangle

Resolution for ISAR and Spotlight SAR

To determine the resolution of an ISAR image, the relationship between the image and the target is analyzed.

For turntable geometry, (30) is used. The viewing direction is taken to be $\boldsymbol{x}^0 = (1,0,0)$, with $\hat{\boldsymbol{e}}_\theta = \mathcal{O}(\theta)\boldsymbol{x}^0$ and $\tilde{k} = 2k$. Then (32) is proportional to

$$\begin{aligned}\widetilde{D}(\tilde{k},\theta) &= \int \mathrm{e}^{-\mathrm{i}\tilde{k}\hat{\boldsymbol{e}}_\theta\cdot\boldsymbol{y}} q(\mathbf{y})\,\mathrm{d}\mathbf{y} \\ &= \iint \mathrm{e}^{-\mathrm{i}\tilde{k}(y_1\cos\theta + y_2\sin\theta)} \underbrace{\int q(y_1,y_2,y_3)\,\mathrm{d}y_3}_{\tilde{q}(y_1,y_2)}\,\mathrm{d}y_1\,\mathrm{d}y_2. \end{aligned} \tag{48}$$

The data depend only on the quantity $\tilde{q}(y_1,y_2) = \int q(y_1,y_2,y_3)\,\mathrm{d}y_3$, which is a projection of the target onto the plane orthogonal to the axis of rotation. In other words, in the turntable geometry, the radar imaging projection is the projection onto the horizontal plane. With the notation $\mathbf{y} = (y_1,y_2)$, so that $\boldsymbol{y} = (\mathbf{y}, y_3)$, it is clear that $\hat{\boldsymbol{e}}_\theta\cdot\boldsymbol{y} = (\mathcal{P}\hat{\boldsymbol{e}}_\theta)\cdot\mathbf{y}$, where $\mathcal{P}:\mathbb{R}^3\to\mathbb{R}^2$ denotes the projection onto the first two components of a three-dimensional vector.

The data-collection manifold $\Omega = \{\tilde{k}\hat{\boldsymbol{e}}_\theta : \omega_1 < \omega < \omega_2 \text{ and } |\theta| < \Phi\}$ is shown in Fig. 2. Then (48) can be written as

$$\widetilde{D}(\tilde{k},\theta) = \chi_\Omega(\tilde{k}\hat{\boldsymbol{e}}_\theta)\mathcal{F}[\tilde{q}](\tilde{k}\hat{\boldsymbol{e}}_\theta), \tag{49}$$

where $\chi_\Omega(\tilde{k}\hat{\boldsymbol{e}}_\theta)$ denotes the function that is 1 if $\tilde{k}\hat{\boldsymbol{e}}_\theta \in \Omega$ and 0 otherwise.

The image is formed by taking the two-dimensional inverse Fourier transform of (49):

$$\begin{aligned} I(\mathbf{x}) &= \iint \mathrm{e}^{\mathrm{i}\mathbf{x}\cdot\tilde{k}(\mathcal{P}\hat{\boldsymbol{e}}_\theta)}\widetilde{D}(\tilde{k},\theta)\tilde{k}\,\mathrm{d}\tilde{k}\,d\theta \propto \int_\Omega e^{i\mathbf{x}\cdot\tilde{k}(\mathcal{P}\hat{\boldsymbol{e}}_\theta)} \iint e^{-i\mathbf{y}\cdot\tilde{k}(\mathcal{P}\hat{\boldsymbol{e}}_\theta)}\tilde{q}(\mathbf{y})\,\mathrm{d}\mathbf{y}\tilde{k}\,\mathrm{d}\tilde{k}\;\mathrm{d}\theta \\ &= \int \underbrace{\iint_\Omega \mathrm{e}^{\mathrm{i}(\mathbf{x}-\mathbf{y})\cdot\tilde{k}(\mathcal{P}\hat{\boldsymbol{e}}_\theta)}\tilde{k}\,\mathrm{d}\tilde{k}\,\mathrm{d}\theta}_{K(\mathbf{x}-\mathbf{y})}\,\tilde{q}(\mathbf{y})\,\mathrm{d}\mathbf{y}. \end{aligned} \tag{50}$$

The function K is the *point-spread function* (PSF); it is also called the *imaging kernel*, *impulse response*, or sometimes *ambiguity function*. The PSF can be written as

$$K(\mathbf{x}) \propto \iint_\Omega \mathrm{e}^{\mathrm{i}\mathbf{x}\cdot\tilde{k}(\mathcal{P}\hat{\boldsymbol{e}}_\theta)}\tilde{k}\,\mathrm{d}\tilde{k}\,\mathrm{d}\theta = \int_{|\xi|=\tilde{k}_1}^{|\xi|=\tilde{k}_2}\int_{-\Phi}^{\Phi} \mathrm{e}^{\mathrm{i}\mathbf{x}\cdot\tilde{k}(\mathcal{P}\hat{\boldsymbol{e}}_\theta)}\tilde{k}\,\mathrm{d}\tilde{k}\,\mathrm{d}\theta. \tag{51}$$

It can be calculated by writing

$$\mathbf{x} = r(\cos\psi, \sin\psi) \quad \text{and} \quad (\mathcal{P}\hat{\boldsymbol{e}}_\theta) = (\cos\phi, \sin\phi), \tag{52}$$

so that $\mathbf{x} \cdot (\mathcal{P}\hat{\boldsymbol{e}}_\theta) = r\cos(\phi - \psi)$. The "down-range" direction corresponds to $\psi = 0$ and "cross-range" corresponds to $\psi = \pi/2$.

Down-Range Resolution in the Small-Angle Case

For many radar applications, the target is viewed from only a small range of aspects $\hat{\boldsymbol{e}}_\theta$; in this case, the small-angle approximations $\cos\phi \approx 1$ and $\sin\phi \approx \phi$ can be used.

In the down-range direction ($\psi = 0$), under the small-angle approximation, (51) becomes

$$\begin{aligned} K(r,0) &\approx \int_{\tilde{k}_1}^{\tilde{k}_2} \tilde{k} \int_{-\Phi}^{\Phi} \mathrm{e}^{\mathrm{i}\tilde{k}r}\, \mathrm{d}\phi\, \mathrm{d}\tilde{k} \\ &= 2\Phi \int_{\tilde{k}_1}^{\tilde{k}_2} \tilde{k}\mathrm{e}^{\mathrm{i}\tilde{k}r}\, \mathrm{d}\tilde{k} = \frac{2\Phi}{\mathrm{i}} \frac{\mathrm{d}}{\mathrm{d}r} \int_{\tilde{k}_1}^{\tilde{k}_2} \mathrm{e}^{\mathrm{i}\tilde{k}r}\, \mathrm{d}\tilde{k} \\ &= \frac{2\Phi}{\mathrm{i}} \frac{\mathrm{d}}{\mathrm{d}r} \left[\mathrm{e}^{\mathrm{i}\tilde{k}_0 r} \frac{b}{2} \operatorname{sinc} \frac{br}{2} \right], \end{aligned} \tag{53}$$

where $b = \tilde{k}_2 - \tilde{k}_1 = 4\pi B/c$, B is the bandwidth in hertz, and $\tilde{k}_0 = (\tilde{k}_1 + \tilde{k}_2)/2 = 2\pi(\nu_1 + \nu_2) = 2\pi\nu_0$, where ν_0 is the center frequency in hertz.

Since $\tilde{k}_0 \gg b$, the leading order term of (53) is obtained by differentiating the exponential:

$$K(r,0) \approx b\tilde{k}_0 \Phi\, \mathrm{e}^{\mathrm{i}\tilde{k}_0 r} \operatorname{sinc} \frac{1}{2} br, \tag{54}$$

yielding peak-to-null down-range resolution $2\pi/b = c/(2B)$. Here, it is the sinc function that governs the resolution.

Cross-Range Resolution in the Small-Angle Case

In the cross-range direction ($\psi = \pi/2$), the approximation $\cos(\phi - \psi) = \sin\phi \approx \phi$ holds under the small-angle assumption. With this approximation, the computation of (51) is

$$\begin{aligned} K(0,r) &\approx \int_{\tilde{k}_1}^{\tilde{k}_2} \tilde{k} \int_{-\Phi}^{\Phi} \mathrm{e}^{\mathrm{i}\tilde{k}r\phi}\, \mathrm{d}\phi\, \mathrm{d}\tilde{k} \\ &= \int_{\tilde{k}_1}^{\tilde{k}_2} \tilde{k} \frac{\mathrm{e}^{\mathrm{i}\tilde{k}r\Phi} - \mathrm{e}^{-\mathrm{i}\tilde{k}r\Phi}}{\mathrm{i}\tilde{k}r}\, \mathrm{d}\tilde{k} \\ &= \frac{1}{\mathrm{i}r} \left[\mathrm{e}^{\mathrm{i}\tilde{k}_0 r\Phi} b \operatorname{sinc}\left(\frac{1}{2} br\Phi\right) - \mathrm{e}^{-\mathrm{i}\tilde{k}_0 r\Phi} b \operatorname{sinc}\left(\frac{1}{2} br\Phi\right) \right] \\ &= 2b\tilde{k}_0 \Phi \operatorname{sinc}\left(\frac{1}{2} br\Phi\right) \operatorname{sinc}(\tilde{k}_0 r\Phi). \end{aligned} \tag{55}$$

Since $\tilde{k}_0 \gg b$,

$$K(0,r) \approx 2b\tilde{k}_0\Phi\,\mathrm{sinc}(\tilde{k}_0 r\Phi). \tag{56}$$

Thus the peak-to-null cross-range resolution is $\pi/(\tilde{k}_0\Phi) = c/(4\nu_0\Phi) = \lambda_0/(4\Phi)$. Since the angular aperture is 2Φ, the cross-range resolution is λ_0 divided by twice the angular aperture.

Example Figure 9 shows a numerical calculation of K for $\phi = 12°$ and two different frequency bands: $[\tilde{k}_1, \tilde{k}_2] = [0, 300]$ (i.e., $b = 300$ and $\tilde{k}_0 = 150$, and $[\tilde{k}_1, \tilde{k}_2] = [200, 300]$) (i.e., $b = 100$ and $\tilde{k}_0 = 250$). The first case is not relevant for most radar systems, which do not transmit frequencies near zero, but is relevant for other imaging systems such as X-ray tomography. These results are plotted in Fig. 9.

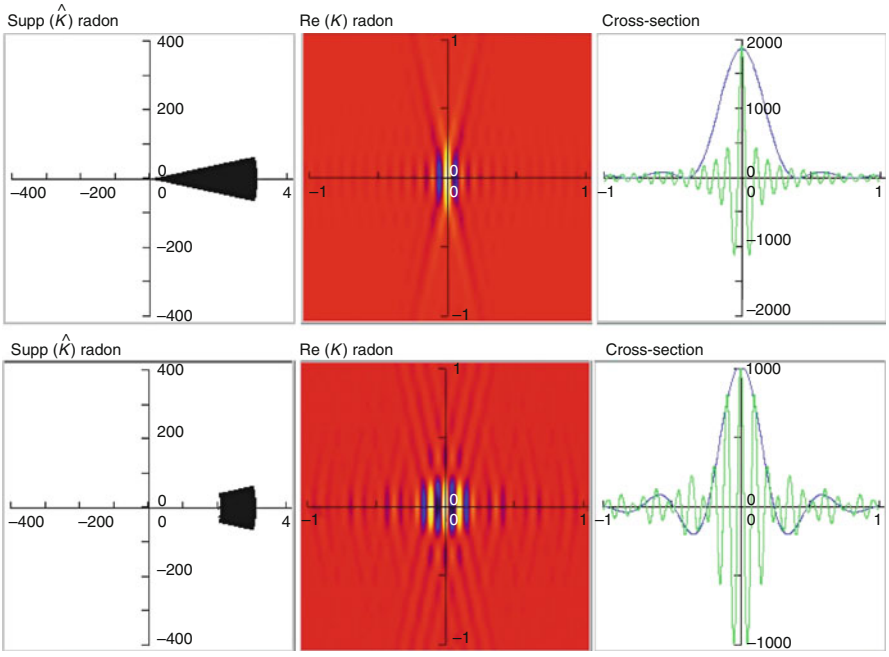


Fig. 9 From *left* to *right*: the data-collection manifold, the real part of K, cross sections (horizontal is rapidly oscillating; vertical is slowly oscillating) through the real part of K for the two cases. Down range is horizontal (Reprinted with permission from [45])

5 Numerical Methods

ISAR and Spotlight SAR Algorithms

The Polar Format Algorithm (PFA) For narrow-aperture, turntable-geometry data, such as shown in Fig. 2, the polar format algorithm (PFA) is commonly used. The PFA consists of the following steps, applied to frequency-domain data:

1. Interpolate from a polar grid to a rectangular grid (see Fig. 10.)
2. Use the two-dimensional discrete (inverse) Fourier transform to form an image of q.

Alternatively, algorithms for computing the Fourier transform directly from a nonuniform grid can be used [33, 46, 57].

Inversion by Filtered Backprojection For the n-dimensional Radon transform, one of the many inversion formulas [43, 44] is

$$f = \frac{1}{2(2\pi)^{n-1}} \mathcal{R}^{\dagger} \mathcal{I}^{1-n} \left(\mathcal{R}[f] \right), \tag{57}$$

where $\mathcal{I}^{1-n}$ is the Riesz operator (filter)

$$\mathcal{I}^{\alpha} f = \mathcal{F}^{-1} \left[|\nu|^{-\alpha} \mathcal{F} f \right] \tag{58}$$

operating on the s variable, and the operator $\mathcal{R}^{\dagger}$ is the formal adjoint of $\mathcal{R}$. (Here the term "formal" means that the convergence of the integrals is not considered; the

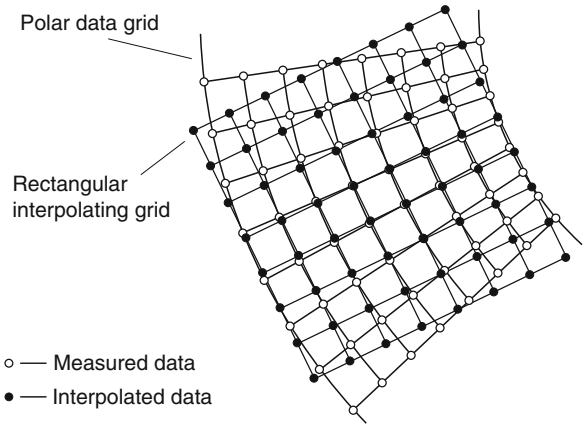


Fig. 10 This illustrates the process of interpolating from a polar grid to a rectangular grid

identities are applied only to functions that decay sufficiently rapidly, so that the integrals converge.) The adjoint is defined by the relation

$$\langle \mathcal{R}f, h\rangle_{s,\widehat{\boldsymbol{\mu}}} = \langle f, \mathcal{R}^\dagger h\rangle_{\boldsymbol{x}}, \tag{59}$$

where

$$\langle \mathcal{R}f, h\rangle_{s,\widehat{\boldsymbol{\mu}}} = \iint \mathcal{R}(s,\widehat{\boldsymbol{\mu}})h^*(s,\widehat{\boldsymbol{u}})\,\mathrm{d}s\,\mathrm{d}\widehat{\boldsymbol{\mu}} \tag{60}$$

and

$$\langle f, \mathcal{R}^\dagger h\rangle_{\boldsymbol{x}} = \int f(\boldsymbol{x})[\mathcal{R}^\dagger h]^*(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}.$$

Using (37) in (60) and interchanging the order of integration show that the adjoint $\mathcal{R}^\dagger$ operates on $h(s,\boldsymbol{\mu})$ via

$$(R^\dagger h)(\boldsymbol{x}) = \int_{S^{n-1}} h(\boldsymbol{x}\cdot\widehat{\boldsymbol{\mu}},\widehat{\boldsymbol{\mu}})\,\mathrm{d}\widehat{\boldsymbol{\mu}}. \tag{61}$$

Here $\mathcal{R}^\dagger$ integrates over the part of h corresponding to all planes ($n = 3$) or lines ($n = 2$) through $\boldsymbol{x}$. When $\mathcal{R}^\dagger$ operates on Radon data, it has the physical interpretation of *backprojection*. For example, in the case where h represents Radon data from a point-like target, for a fixed direction $\hat{\boldsymbol{\mu}}$, the quantity $h(\boldsymbol{x}\cdot\widehat{\boldsymbol{\mu}},\widehat{\boldsymbol{\mu}})$, as a function of $\boldsymbol{x}$, is constant along each plane (or line if $n = 2$) $\boldsymbol{x}\cdot\hat{\boldsymbol{\mu}} =$ constant. Thus, at each $\hat{\boldsymbol{\mu}}$, the function $h(\boldsymbol{x}\cdot\widehat{\boldsymbol{\mu}},\widehat{\boldsymbol{\mu}})$ can be thought of as an image in which the data h for direction $\hat{\boldsymbol{u}}$ is backprojected (smeared) onto all points $\boldsymbol{x}$ that could have produced the data for that direction. The integral in (61) then sums the contributions from all the possible directions. (See Fig. 11.) The inversion formula (57) is thus a *filtered backprojection* formula. Fast backprojection algorithms have been developed by a number of authors (e.g., [24, 76]).

Range Alignment

ISAR imaging relies on target/radar relative motion. An assumption made throughout is that the target moves as a rigid body – an assumption that ignores the flexing of aircraft lift and control surfaces or the motion of vehicle treads. Moreover, arbitrary rigid-body motion can always be separated into a rotation about the body's center of mass and a translation of that center of mass. Backprojection shows how the rotation part of the relative radar/target motion can be used to reconstruct a two-dimensional image of the target in ISAR and spotlight SAR. But, usually while the target is rotating and the radar system is collecting data, *the target will also be translating*,

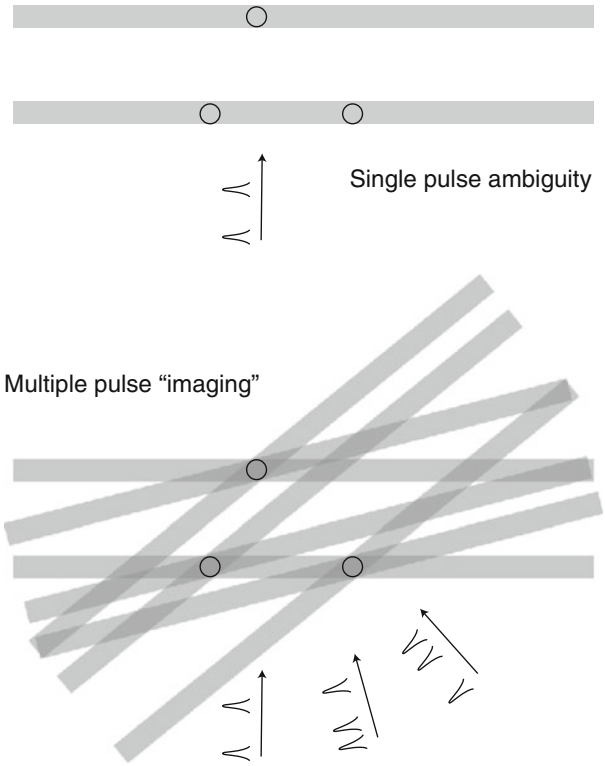


Fig. 11 This figure illustrates the process of backprojection. The range profiles (*inset*) suggest the time delays that would result from an interrogating radar pulse incident from the indicated direction. Note that scatterers that lie at the same range from one view do *not* lie at the same range from other views

and this has not been accounted for in the imaging algorithm. In Eqs. (27) and (28), for example, $R = |\boldsymbol{x}_0|$ was implicitly set to a constant (Fig. 12).

Typically, the radar data are preprocessed to subtract out the effects of target translation before the imaging step is performed. Under the start–stop approximation, the range profile data $\eta_B(t, \theta_n)$ is approximately a shifted version of the range profile (see section "The Range Profile") at the previous pulse. Thus $\eta_B(t, \theta_{n+1}) \approx \eta_B(t + \Delta t_n, \theta_n)$, where Δt_n is a range offset that is determined by target motion between pulses.

The collected range profiles can be shifted to a common origin if Δt_n can be determined for each θ_n. One method to accomplish this is to assume that one of the peaks in each of the range profiles (e.g., the strongest peak) is always due to the *same* target feature and so provides a convenient origin. This correction method is known as "range alignment" and must be very accurate in order to correct the offset error within a fraction of a wavelength. (Note that the wavelength in question is that of the signal output by the correlation receiver and not the wavelength of the

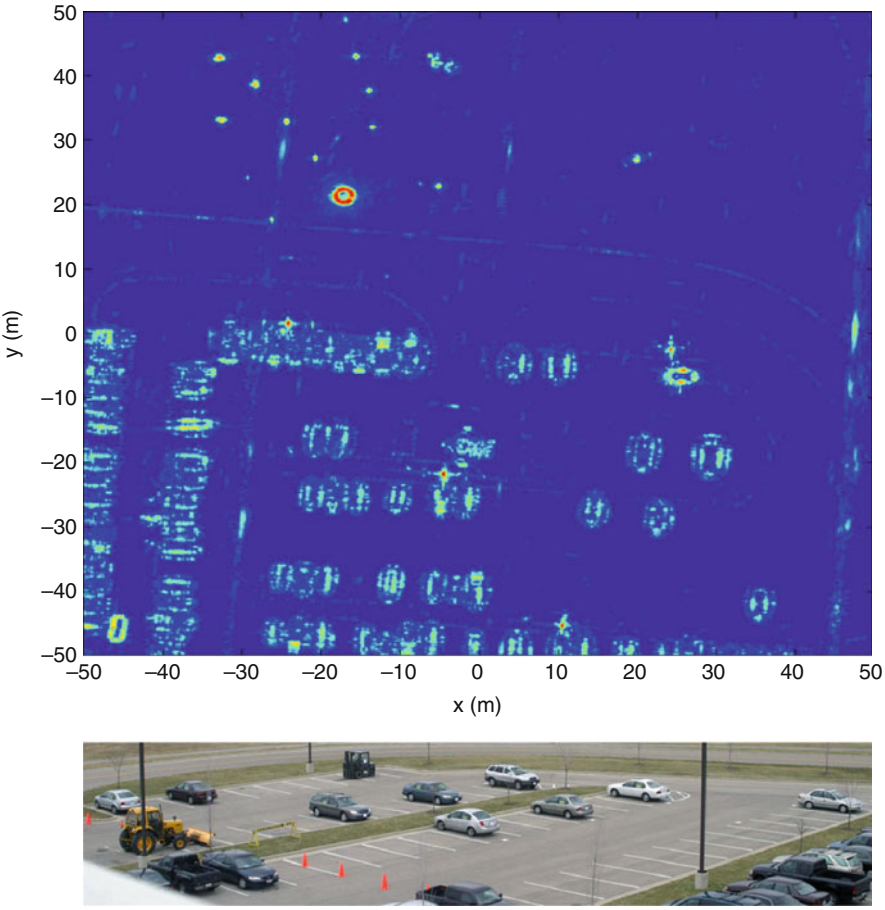


Fig. 12 A radar image from a circular flight path, together with an optical image of the same scene. The bright ring in the top half of the radar image is a "top-hat" calibration target used to focus the image (Courtesy US Air Force Sensors Directorate)

transmitted waveform. In HRR systems, however, this wavelength can still be quite small.) Typically, $\eta_B(t + \Delta t_n, \theta_n)$ is correlated with $\eta_B(t, \theta_{n+1})$, and the correlation maximum is taken to indicate the size of the shift Δt_n. This idea is illustrated in Fig. 13 which displays a collection of properly aligned range profiles.

When the scattering center used for range alignment is not a single point but, rather, several closely spaced and *unresolved* scattering centers, then additional constructive and destructive interference effects can cause the range profile alignment feature – assumed to be due to a single well-localized scatterer – to vary rapidly across the synthetic aperture (i.e., such scattering centers are said to "scintillate"). For very complex and scintillating targets, other alignment methods are used: for example, if the target is assumed to move along a "smooth" path, then estimates

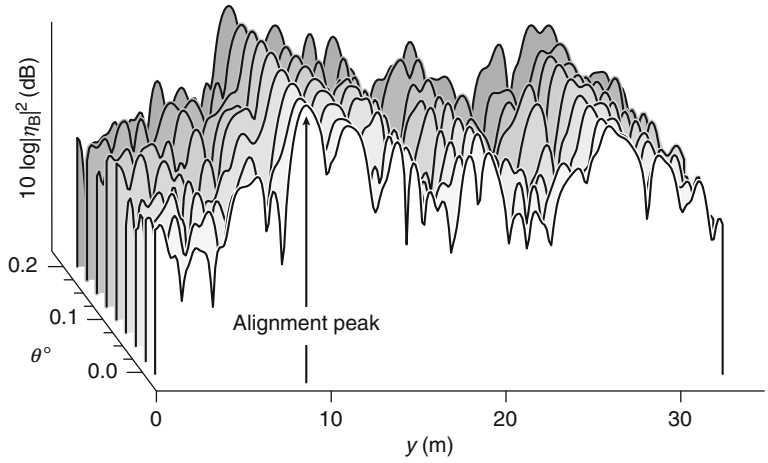


Fig. 13 Range alignment preprocessing in synthetic aperture imaging. The effects of target translation must be removed before backprojection can be applied

of its range, range rate, range acceleration, and range jerk (time derivative of acceleration) can be used to express target range as a polynomial in time

$$R(\theta_n) = R(0) + \dot{R}\theta_n + \frac{1}{2}\ddot{R}\theta_n^2 + \frac{1}{6}\dddot{R}\theta_n^3. \tag{62}$$

In terms of this polynomial,

$$\Delta t_n = 2\,\frac{R(\theta_n) - R(0)}{c} = 2\,\frac{\dot{R}\theta_n + \frac{1}{2}\ddot{R}\theta_n^2 + \frac{1}{6}\dddot{R}\theta_n^3}{c}, \tag{63}$$

where $\dot{R}$, $\ddot{R}$, and $\dddot{R}$ are radar measurables.

Of course, the need for range alignment preprocessing is not limited to ISAR imaging; similar *motion compensation* techniques are needed in SAR as well (Fig. 14).

6 Open Problems

In the decades since the invention of synthetic aperture radar imaging, there has been much progress, but many open problems still remain. And most of these open problems are mathematical in nature.

As outlined at the beginning of Sect. 3, SAR imaging is based on specific assumptions, which in practice may not be satisfied. When they are not satisfied, artifacts appear in the image.

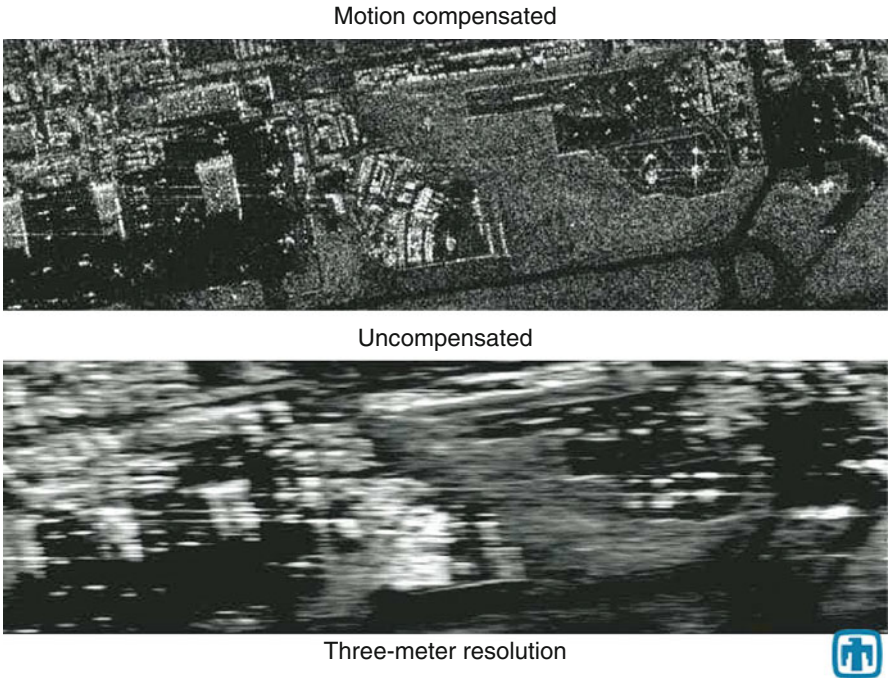


Fig. 14 The effect of motion compensation in a Ku-band image (Courtesy Sandia National Laboratories)

Problems Related to Unmodeled Motion

SAR image formation algorithms assume the scene to be stationary. Motion in the scene gives rise to mispositioning or streaking (see Figs. 15 and 16). This effect is analyzed in [28].

However, it is of great interest to use radar to identify moving objects; systems that can do this are called *moving target indicator* (MTI) systems or *ground moving target indicator* (GMTI) systems.

1. How can artifacts associated with targets that move during data collection [54] be mitigated? Moving targets cause Doppler shifts and also present different aspects to the radar [14]. An approach for exploiting unknown motion is given in [65].
2. Both SAR and ISAR are based on known relative motion between the target and sensor, for example, the assumption that the target behaves as a rigid body. When this is not the case, the images are blurred or uninterpretable. Better methods for finding the relative motion between the target and sensor are also needed [6, 65]. Better algorithms are needed for determining the antenna position from the radar

Fig. 15 A Ku-band image showing streaking due to objects moving in the scene (Courtesy Sandia National Laboratories and SPIE)

data itself. Such methods include *autofocus* algorithms [35, 40], some of which use a criterion such as image contrast to focus the image.

3. When the target motion is complex (pitching, rolling, and yawing), it may be possible to form a three-dimensional image; fast, accurate methods for doing this are needed [65]. How can moving objects be simultaneously tracked [58] and imaged?

Problems Related to Unmodeled Scattering Physics

1. How can images be formed without the Born approximation? The Born approximation leaves out many physical effects, including not only multiple scattering and creeping waves but also shadowing, obscuration, and polarization changes. But without the Born approximation (or the Kirchhoff approximation, which is similar), the imaging problem is nonlinear. In particular, how can images be formed in the presence of multiple scattering? (See [6, 13, 31, 49, 73].) Artifacts due to the Born approximation can be seen in Fig. 4, where the vertical streaks near the tail are due to multiple scattering in the engine inlets. Can multiple scattering be exploited [13, 39] to improve resolution?

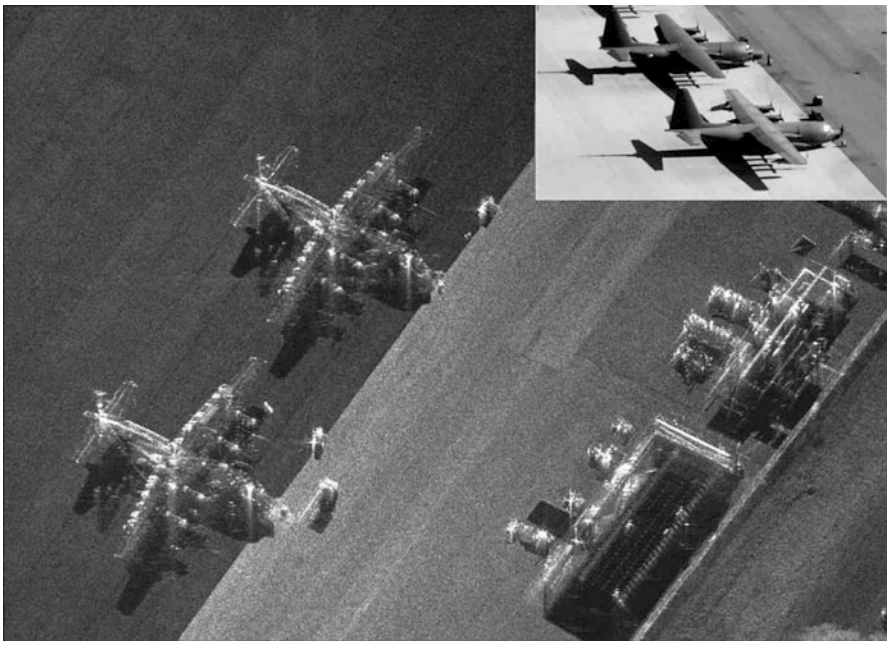


Fig. 16 A 4-in. resolution of SAR image from Sandia National Laboratories. Only certain parts of the airplanes reflect radar energy. The *inset* is an optical image of the airplanes (Courtesy Sandia National Laboratories)

2. Scattering models need to be developed that include as much of the physics as possible but are still simple enough for use in the inverse problem. An example of a simple model that includes relevant physics is [56].
3. How can polarization information [4, 17, 18, 55, 71] be exploited? This problem is closely connected to the issue of multiple scattering: the usual linear models predict no change in the polarization of the backscattered electric field. Consequently linear imaging methods cannot provide information about how scatterers change the polarization of the interrogating field. A paper that may be useful here is [68].
4. How can prior knowledge about the scene be incorporated in order to improve resolution? There is interest in going beyond simple aperture/bandwidth-defined resolution [45,66]. One approach that has been suggested is to apply compressive sensing ideas [1, 10, 42] to SAR.
5. How can information in the radar shadow be exploited? In many cases it is easier to identify an object from its shadow than from its direct-scattering image. (See Fig. 17.) A backprojection method for reconstructing an object's three-dimensional shape from its shadows obtained at different viewing angles is proposed in [23]. What determines the resolution of this reconstruction?

Fig. 17 A 4-in. resolution image from Sandia National Laboratories. Note the shadows of the historical airplane, helicopter, and trees (Courtesy of Sandia National Laboratories)

New Applications of Radar Imaging

1. Can radar systems be used to identify individuals by their gestures or gait? Time-frequency analysis of radar signals gives rise to *micro-Doppler* time-frequency images [11], in which the motion of arms and legs can be identified.
2. How can radar be used to form images of urban areas? It is difficult to form SAR images of urban areas, because in cities the waves undergo complicated multipath scattering. Areas behind buildings lie in the radar shadows, and images of tall buildings can obscure other features of interest. In addition, urban areas tend to be sources of electromagnetic radiation that can interfere with the radiation used for imaging.

   One approach that is being explored is to use a *persistent* or *staring* radar system [27] that would fly in circles [61] around a city of interest (see, e.g., Fig. 12). Thus, the radar would eventually illuminate most of the areas that would be shadowed when viewed from a single direction. However, this approach has the added difficulty that the same object will look different when viewed from different directions. How can the data from a staring radar system be used to obtain the maximum amount of information about the (potentially changing) scene?

3. If sensors are flown on unoccupied aerial vehicles (UAVs), where should these UAVs fly? The notion of swarms of UAVs [2] gives rise not only to challenging problems in control theory but also to challenging imaging problems.
4. Many of these problems motivate a variety of more theoretical open problems such as the question of whether backscattered data uniquely determines a penetrable object or a non-convex surface [63, 72]. There is a close connection between radar imaging and the theory of Fourier integral operators [48]. How can this theory be extended to the case of dispersive media and to nonlinear operators? Is it possible to develop a theory of the information content [37, 52] of an imaging system?

7 Conclusion

Radar imaging is a mathematically rich field with many interesting open problems.

Acknowledgments The authors would like to thank the Naval Postgraduate School and the Air Force Office of Scientific Research which supported the writing of this article under agreement number FA9550-09-1-0013 (because of this support, the US Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Air Force Research Laboratory or the US Government).

Cross-References

▸ Inverse Scattering
▸ Linear Inverse Problems
▸ Tomography
▸ Wave Phenomena

References

1. Baraniuk, R., Steeghs, P.: Compressive radar imaging. In: IEEE Radar Conference, Waltham (2007)
2. Bethke, B., Valenti, M., How, J.P., Vian, J.: Cooperative vision based estimation and tracking using multiple UAVs. In: Conference on Cooperative Control and Optimization, Gainesville (2007)
3. Bleistein, N., Cohen, J.K., Stockwell, J.W.: The Mathematics of Multidimensional Seismic Inversion. Springer, New York (2000)
4. Boerner, W.-M., Yamaguchi, Y.: A state-of-the-art review in radar polarimetry and its applications in remote sensing. IEEE Aerosp. Electron. Syst. Mag. **5**, 3–6 (1990)
5. Borden, B.: Radar Imaging of Airborne Targets. Institute of Physics, Bristol (1999)

6. Borden, B.: Mathematical problems in radar inverse scattering. Inverse Probl. **18**, R1–R28 (2002)
7. Bowen, E.G.: Radar Days. Hilgar, Bristol (1987)
8. Buderi, R.: The Invention That Changed the World. Simon & Schuster, New York (1996)
9. Carrara, W.C., Goodman, R.G., Majewski, R.M.: Spotlight Synthetic Aperture Radar: Signal Processing Algorithms. Artech House, Boston (1995)
10. Cetin, M., Karl, W.C., Castañon, D.A.: Analysis of the impact of feature-enhanced SAR imaging on ATR performance. In: Algorithms for SAR Imagery IX, Proceedings of SPIE, vol. 4727 (2002)
11. Chen, V.C., Ling, H.: Time-Frequency Transforms for Radar Imaging and Signal Analysis. Artech House, Boston (2002)
12. Cheney, M.: A mathematical tutorial on synthetic aperture radar. SIAM Rev. **43**, 301–312 (2001)
13. Cheney, M., Bonneau, R.J.: Imaging that exploits multipath scattering from point scatterers. Inverse Probl. **20**, 1691–1711 (2004)
14. Cheney, M., Borden, B.: Imaging moving targets from scattered waves. Inverse Probl. **24**, 035005 (2008)
15. Cheney, M., Borden, B.: Fundamentals of Radar Imaging. SIAM, Philadelphia (2009)
16. Chew, W.C., Song, J.M.: Fast Fourier transform of sparse spatial data to sparse Fourier data. In: IEEE Antenna and Propagation International Symposium, vol. 4, pp. 2324–2327 (2000)
17. Cloude, S.R.: Polarization coherence tomography. Radio Sci. **41**, RS4017 (2006). doi:10.1029/2005RS003436
18. Cloude, S.R., Papathanassiou, K.P.: Polarimetric SAR interferometry. IEEE Trans. Geosci. Remote Sens. **36**(5, part 1), 1551–1565 (1998)
19. Cook, C.E., Bernfeld, M.: Radar Signals. Academic, New York (1967)
20. Cumming, I.G., Wong, F.H.: Digital Processing of Synthetic Aperture Radar Data: Algorithms and Implementation. Artech House, Boston (2005)
21. Curlander, J.C., McDonough, R.N.: Synthetic Aperture Radar. Wiley, New York (1991)
22. Cutrona, L.J.: Synthetic aperture radar. In: Skolnik, M. (ed.) Radar Handbook, 2nd edn. McGraw-Hill, New York (1990)
23. Dickey, F.M., Doerry, A.W.: Recovering shape from shadows in synthetic aperture radar imagery. In: Ranney, K.I., Doerry, A.W. (eds.) Radar Sensor Technology XII. Proceedings of SPIE, vol. 6947, pp. 694707 (2008)
24. Ding, Y., Munson, D.C. Jr.: A fast back-projection algorithm for bistatic SAR imaging. In: Proceedings of the IEEE International Conference on Image Processing, Rochester, 22–25 Sept 2002 (2002)
25. Edde, B.: Radar: Principles, Technology, Applications. Prentice-Hall, Englewood Cliffs (1993)
26. Elachi, C.: Spaceborne Radar Remote Sensing: Applications and Techniques. IEEE, New York (1987)
27. Ertin, E., Austin, C.D., Sharma, S., Moses, R.L., Potter, L.C.: GOTCHA experience report: three-dimensional SAR imaging with complete circular apertures. Proc. SPIE **6568**, 656802 (2007)
28. Fienup, J.R.: Detecting moving targets in SAR imagery by focusing. IEEE Trans. Aerosp. Electron. Syst. **37**, 794–809 (2001)
29. Franceschetti, G., Lanari, R.: Synthetic Aperture Radar Processing. CRC, New York (1999)
30. Friedlander, F.G.: Introduction to the Theory of Distributions. Cambridge University Press, New York (1982)
31. Garnier, J., Sølna, K.: Coherent interferometric imaging for synthetic aperture radar in the presence of noise. Inverse Probl. **24**, 055001 (2008)
32. Giuli, D.: Polarization diversity in radars. Proc. IEEE **74**(2), 245–269 (1986)
33. Greengard, L., Lee, J.-Y.: Accelerating the nonuniform fast Fourier transform. SIAM Rev. **46**, 443–454 (2004)
34. Jackson, J.D.: Classical Electrodynamics, 2nd edn. Wiley, New York (1962)
35. Jakowatz, C.V., Wahl, D.E., Eichel, P.H., Ghiglia, D.C., Thompson, P.A.: Spotlight-Mode Synthetic Aperture Radar: A Signal Processing Approach. Kluwer, Boston (1996)

36. Ishimaru, A.: Wave Propagation and Scattering in Random Media. IEEE, New York (1997)
37. Klug, A., Crowther, R.A.: Three-dimensional image reconstruction from the viewpoint of information theory. Nature **238**, 435–440 (1972). doi:10.1038/238435a0
38. Langenberg, K.J., Brandfass, M., Mayer, K., Kreutter, T., Brüll, A., Felinger, P., Huo, D.: Principles of microwave imaging and inverse scattering. EARSeL Adv. Remote Sens. **2**, 163–186 (1993)
39. Lerosey, G., de Rosny, J., Tourin, A., Fink, M.: Focusing beyond the diffraction limit with far-field time reversal. Science **315**, 1120–1122 (2007)
40. Lee-Elkin, F.: Autofocus for 3D imaging. Proc. SPIE **6970**, 69700O (2008)
41. Mensa, D.L.: High Resolution Radar Imaging. Artech House, Dedham (1981)
42. Moses, R., Çetin, M., Potter, L.: Wide Angle SAR Imaging (SPIE Algorithms for Synthetic Aperture Radar Imagery XI). SPIE, Orlando (2004)
43. Natterer, F.: The Mathematics of Computerized Tomography. SIAM, Philadelphia (2001)
44. Natterer, F., Wübbeling, F.: Mathematical Methods in Imaging Reconstruction. SIAM, Philadelphia (2001)
45. Natterer, F., Cheney, M., Borden, B.: Resolution for radar and X-ray tomography. Inverse Probl. **19**, S55–S64 (2003)
46. Nguyen, N., Liu, Q.H.: The regular Fourier matrices and nonuniform fast Fourier transforms. SIAM J. Sci. Comput. **21**, 283–293 (1999)
47. Newton, R.G.: Scattering Theory of Waves and Particles. Dover, Mineola (2002)
48. Nolan, C.J., Cheney, M.: Synthetic aperture inversion for arbitrary flight paths and non-flat topography. IEEE Trans. Image Process. **12**, 1035–1043 (2003)
49. Nolan, C.J., Cheney, M., Dowling, T., Gaburro, R.: Enhanced angular resolution from multiply scattered waves. Inverse Probl. **22**, 1817–1834 (2006)
50. North, D.O.: Analysis of the factors which determine signal/noise discrimination in radar. Report PPR 6C, RCA Laboratories, Princeton (classified) (1943). Reproduction: North, D.O.: An analysis of the factors which determine signal/noise discrimination in pulsed carrier systems. Proc. IEEE **51**(7), 1016–1027 (1963)
51. Oppenheim, A.V., Shafer, R.W.: Digital Signal Processing. Prentice-Hall, Englewood Cliffs (1975)
52. O'Sullivan, J.A., Blahut, R.E., Snyder, D.L.: Information-theoretic image formation. IEEE Trans. Inf. Theory **44**, 2094–2123 (1998)
53. Oughstun, K.E., Sherman, G.C.: Electromagnetic Pulse Propagation in Causal Dielectrics. Springer, New York (1997)
54. Perry, R.P., DiPietro, R.C., Fante, R.L.: SAR imaging of moving targets. IEEE Trans Aerosp. Electron. Syst. **35**(1), 188–200 (1999)
55. Pike, R., Sabatier, P.: Scattering: Scattering and Inverse Scattering in Pure and Applied Science. Academic, New York (2002)
56. Potter, L.C., Moses, R.L.: Attributed scattering centers for SAR ATR. IEEE Trans. Image Process. **6**, 79–91 (1997)
57. Potts, D., Steidl, G., Tasche, M.: Fast Fourier transforms for nonequispaced data: a tutorial. In: Benedetto, J.J., Ferreira, P. (eds.) Modern Sampling Theory: Mathematics and Applications, chap. 12, pp. 249–274. Birkhäuser, Boston (2001)
58. Ramachandra, K.V.: Kalman Filtering Techniques for Radar Tracking. CRC, Boca Raton (2000)
59. Rihaczek, A.W.: Principles of High-Resolution Radar. McGraw-Hill, New York (1969)
60. Skolnik, M.: Introduction to Radar Systems. McGraw-Hill, New York (1980)
61. Soumekh, M.: Synthetic Aperture Radar Signal Processing with MATLAB Algorithms. Wiley, New York (1999)
62. Stakgold, I.: Green's Functions and Boundary Value Problems, 2nd edn. Wiley-Interscience, New York (1997)
63. Stefanov, P., Uhlmann, G.: Inverse backscattering for the acoustic equation. SIAM J. Math. Anal. **28**, 1191–1204 (1997)
64. Stimson, G.W.: Introduction to Airborne Radar. SciTech, Mendham (1998)

65. Stuff, M.A., Sanchez, P., Biancala, M.: Extraction of three-dimensional motion and geometric invariants. Multidimens. Syst. Signal Process. **14**, 161–181 (2003)
66. Sullivan, R.J.: Radar Foundations for Imaging and Advanced Concepts. SciTech, Raleigh (2004)
67. Swords, S.S.: Technical History of the Beginnings of Radar. Peregrinus, London (1986)
68. Treuhaft, R.N., Siqueira, P.R.: Vertical structure of vegetated land surfaces from interferometric and polarimetric radar. Radio Sci. **35**(1), 141–177 (2000)
69. Treves, F.: Basic Linear Partial Differential Equations. Academic, New York (1975)
70. Trischman, J.A., Jones, S., Bloomfield, R., Nelson, E., Dinger, R.: An X-band linear frequency modulated radar for dynamic aircraft measurement. In: AMTA Proceedings, p. 431. AMTA, New York (1994)
71. Ulaby, F.T., Elachi, C. (eds.): Radar Polarimetry for Geoscience Applications. Artech House, Norwood
72. Walsh, T.E.: Military radar systems: history, current position, and future forecast. Microw. J **21**, 87, 88, 91–95 (1978)
73. Weglein, A.B., Araùjo, F.V., Carvalho, P.M., Stolt, R.H., Matson, K.H., Coates, R.T., Corrigan, D., Foster, D.J., Shaw, S.A., Zhang, H.: Inverse scattering series and seismic exploration. Inverse Probl. **19**, R27–R83 (2003). doi:10.1088/0266-5611/19/6/R01
74. Wehner, D.: High-Resolution Radar, 2nd edn. Scitech, Raleigh (1995)
75. Woodward, P.M.: Probability and Information Theory, with Applications to Radar. McGraw-Hill, New York (1953)
76. Xiao, S., Munson, D.C., Basu, S., Bresler, Y.: An N2logN back-projection algorithm for SAR image formation. In: Proceedings of 34th Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, 31 Oct–1 Nov 2000

Tomography

Gabor T. Herman

Contents

1 Introduction 802
2 Background 802
3 Mathematical Modeling and Analysis 804
4 Numerical Methods and Case Examples 822
5 Conclusion 842
Cross-References 842
References 842

Abstract

We define *tomography* as the process of producing an image of a distribution (of some physical property) from estimates of its line integrals along a finite number of lines of known locations. We touch upon the computational and mathematical procedures underlying the data collection, image reconstruction, and image display in the practice of tomography. The emphasis is on reconstruction methods, especially the so-called series expansion reconstruction algorithms.

We illustrate the use of tomography (including three-dimensional displays based on reconstructions) both in electron microscopy and in X-ray computerized tomography (CT), but concentrate on the latter. This is followed by a classification and discussion of reconstruction algorithms. In particular, we discuss how to evaluate and compare the practical efficacy of such algorithms.

G.T. Herman (✉)
Department of Computer Science, The Graduate Center of the City University of New York, New York, NY, USA
e-mail: gaborherman@yahoo.com

© Springer Science+Business Media New York 2015

O. Scherzer (ed.), *Handbook of Mathematical Methods in Imaging*,
DOI 10.1007/978-1-4939-0790-8_16

1 Introduction

To get the flavor of tomography in general, we first discuss a special case: X-ray computerized tomography (CT) for reconstructing the distribution within a transverse section of the human body of a physical parameter (the "relative linear attenuation at energy $\bar{e}$" whose value at the point (x, y) in the section is denoted by $\mu_{\bar{e}}(x, y)$) from multiple X-ray projections. A typical method by which data are collected for transverse section imaging in CT is indicated in Fig. 1. A large number of measurements are taken. Each of these measurements is related to an X-ray source position combined with an X-ray detector position, and from the measurements we can (based on physical principles) estimate the line integral of $\mu_{\bar{e}}$ along the line between the source and the detector. The mathematical problem is: given a large number of such projections, reconstruct the image $\mu_{\bar{e}}(x, y)$.

A chapter such as this can only cover a small part of what is known about tomography. A much extended treatment in the same spirit as this chapter is given in [23]. For additional information on mathematical matters related to CT, the reader may consult the books [7, 17, 24, 29, 50]. In particular, because of the mathematical orientation of this handbook, we will not get into the details of the how the line integrals are estimated from the measurements. (Such details can be found in [23]. They are quite complicated: in addition to the *actual measurement* with the patient in the scanner a *calibration measurement* needs to be taken, both of these need to be normalized by the *reference detector* indicated in Fig. 1, correction has to be made for the *beam hardening* that occurs due to the X-ray beam being polychromatic rather than consisting of photons at the desired energy $\bar{e}$, etc.)

2 Background

The problem of image reconstruction from projections has arisen independently in a large number of scientific fields. A most-important version of the problem in medicine is CT; it has revolutionized diagnostic radiology over the past four decades. The 1979 Nobel prize in physiology and medicine was awarded to Allan M. Cormack and Godfrey N. Hounsfield for the development of X-ray computerized tomography [9, 31]. The 1982 Nobel prize in chemistry was awarded to Aaron Klug, one of the pioneers in the use of reconstruction from electron microscopic projections for the purpose of elucidation of biologically important molecular complexes [11, 13]. The 2003 Nobel prize in physiology and medicine was awarded to Paul C. Lauterbur and Peter Mansfield for their discoveries concerning magnetic resonance imaging, which also included the use of image reconstruction from projections methods [38].

In some sense this problem was solved in 1917 by Johann Radon [52]. Let ℓ denote the distance of the line L from the origin, let θ denote the angle made with the x axis by the perpendicular drawn from the origin to L (see Fig. 1), and let $m(\ell, \theta)$ denote the integral of $\mu_{\bar{e}}$ along the line L. Radon proved that

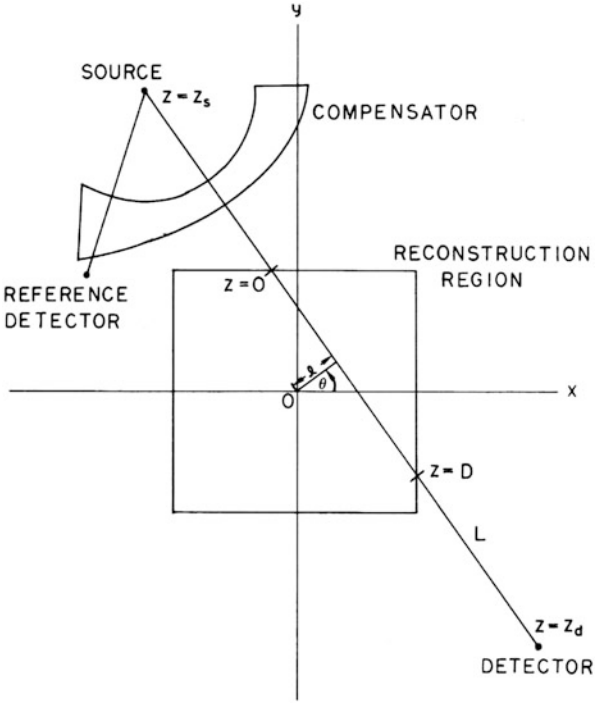


Fig. 1 Data collection for CT (reproduced from [23])

$$\mu_{\tilde{e}}(x, y) = -\frac{1}{2\pi^2} \lim_{\varepsilon \to 0} \int_{\varepsilon}^{\infty} \frac{1}{q} \int_{0}^{2\pi} m_1(x \cos\theta + y \sin\theta + q, \theta)\, d\theta\, dq, \qquad (1)$$

where $m_1(\ell, \theta)$ denotes the partial derivative of $m(\ell, \theta)$ with respect to ℓ. The implication of this formula is clear: the distribution of the relative linear attenuation in an infinitely thin slice is uniquely determined by the set of *all* its line integrals. However:

(a) Radon's formula determines an image from all its line integrals. In CT we have only a finite set of measurements; even if they were *exactly* integrals along lines, a finite number of them would not be enough to determine the image uniquely, or even accurately. Based on the finiteness of the data one can produce objects for which the reconstructions will be very inaccurate (Section 15.4 of [23]).
(b) The measurements in computed tomography can only be used to estimate the line integrals. Inaccuracies in these estimates are due to the width of the X-ray beam, scatter, hardening of the beam, photon statistics, detector inaccuracies, etc. Radon's inversion formula is sensitive to these inaccuracies.
(c) Radon gave a mathematical formula; we need an *efficient* algorithm to evaluate it. This is not necessarily trivial to obtain. There has been a very great deal of

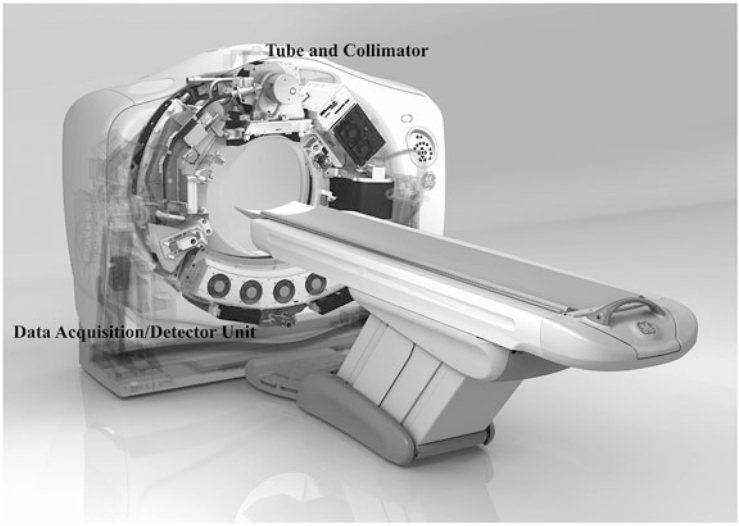


Fig. 2 Engineering rendering of a 2008 CT scanner (provided by GE Healthcare)

activity to find algorithms that are fast when implemented on a computer and yet produce acceptable reconstructions in spite of the finite and inaccurate nature of the data. This chapter concentrates on this topic.

3 Mathematical Modeling and Analysis

The mathematical model for CT is illustrated in Fig. 1. An engineering realization of this model is shown in Fig. 2. The tube contains a single X-ray source, the detector unit contains an array of X-ray detectors. Suppose for the moment that the X-ray Tube and Collimator on the one side and the Data Acquisition/Detector Unit on the other side are stationary, and the patient on the table is moved between them at a steady rate. By shooting a fan beam of X-rays through the patient at frequent regular intervals and detecting them on the other side, we can build up a two-dimensional X-ray projection of the patient that is very similar in appearance to the image that is traditionally captured on an X-ray film. Such a projection is shown in Fig. 3a. The brightness at a point is indicative of the total attenuation of the X-rays from the source to the detector. This mode of operation is *not* CT, it is just an alternative way of taking X-ray images. In the CT mode, the patient is kept stationary, but the tube and the detector unit rotate (together) around the patient. The fan beam of X-rays from the source to the detector determines a slice in the patient's body. The location of such a slice is shown by the horizontal line in Fig. 3a.

Data are collected for a number of fixed positions of the source and detector; these are referred to as *views*. For each view, we have a reading by each of

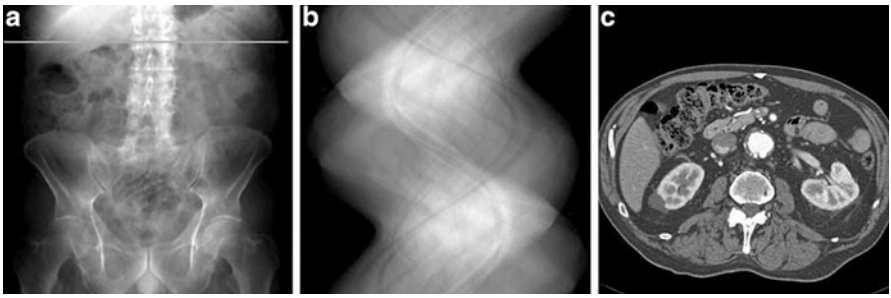


Fig. 3 (**a**) Digitally radiograph with *line* marking the location of the cross section for which the following images were obtained. (**b**) Sinogram of the projection data. (**c**) Reconstruction from the projection data (images were obtained using a Siemens Sensation CT scanner by R. Fahrig and J. Starman at Stanford University)

the detectors. All the detector readings for all the views can be represented as a *sinogram*, shown in Fig. 3b. The intensities in the sinogram are proportional to the line integrals of the X-ray attenuation coefficient between the corresponding source and detector positions. From these line integrals, a two-dimensional image of the X-ray attenuation coefficient distribution in the slice of the body can be produced by the techniques of image reconstruction. Such an image is shown in Fig. 3c. Inasmuch as different tissues have different X-ray attenuation coefficients, boundaries of organs can be delineated and healthy tissue can be distinguished from tumors. In this way CT produces cross-sectional slices of the human body without surgical intervention.

We can use the reconstructions of a series of parallel transverse sections to discover and display the precise shape of selected organs; see Fig. 4. Such displays are obtained by further computer processing of the reconstructed cross sections [59].

As a second illustration of the many applications of tomography (for a more complete coverage see Section 1.1 of [23]), we note that three-dimensional reconstruction of nano-scale objects (such as biological macromolecules) can be accomplished using data recorded with a transmission *electron microscope* (see Fig. 5) that produces *electron micrographs*, such as the one illustrated in Fig. 6, in which the grayness at each point is indicative of a line integral of a physical property of the object being imaged. From multiple electron micrographs one can recover the structure of the object that is being imaged; see Fig. 7.

What we have just illustrated in our electron microscopy example is a reconstruction of a three-dimensional object from two-dimensional projections; as opposed to what is shown in Fig. 1, which describes the collection of data for the reconstruction of a two-dimensional object. In fact, recently developed CT scanners are not like that, they collect a series of two-dimensional projections of the three-dimensional object to be reconstructed.

Helical CT (also referred to as s*piral CT*) first started around 1990 [10, 34] and has become standard for medical diagnostic X-ray CT. Typical state-of-the-

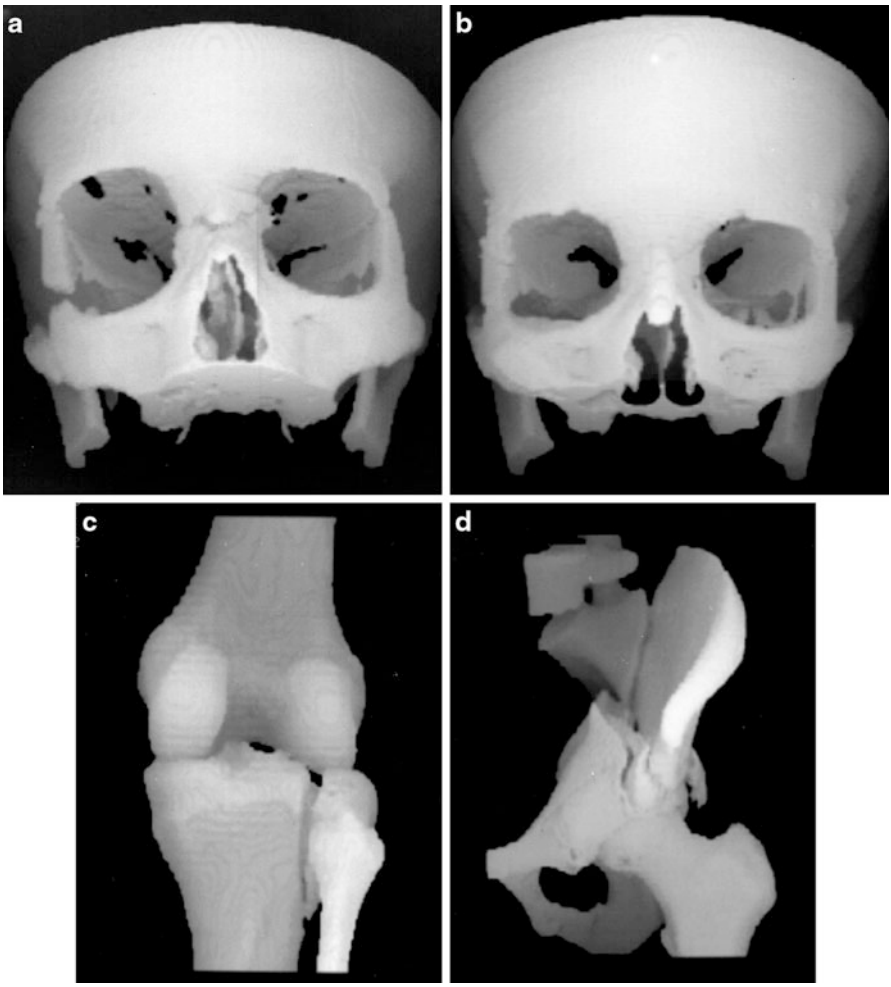


Fig. 4 Three-dimensional displays of bone structures of patients produced during 1986–1988 by software developed in the author's research group at the University of Pennsylvania for the General Electric Company. (**a**) Facial bones of an accident victim prior to operation. (**b**) The same patient at the time of a 1-year postoperative follow-up. (**c**) A tibial fracture. (**d**) A pelvic fracture (reproduced from [23])

art versions of such systems have a single X-ray source and multiple detectors in a two-dimensional array. The main innovation over previously used technologies is the presence of two independent motions: while the source and detectors rotate around the patient, the table on which the patient lies is continuously moved between them (typically orthogonally to the plane of rotation), see Fig. 8. Thus, the trajectory of the source relative to the patient is a helix (hence the name "helical CT"). Helical CT allows rapid imaging as compared with the previous commercially viable approaches, which has potentially many advantages. One example is when

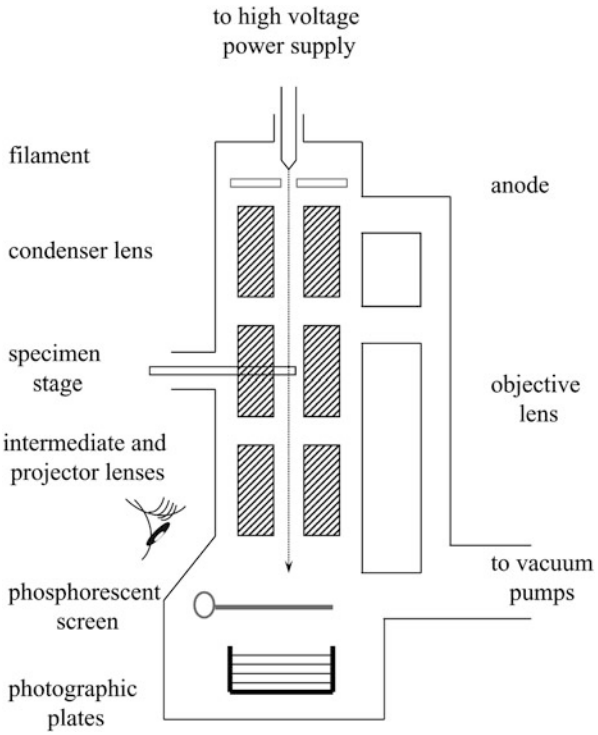


Fig. 5 Schematic drawing of a transmission electron microscope (illustration provided by C. San Martín of the Centro Nacional de Biotecnología, Spain)

we wish to image a long blood vessel that is made visible to X-rays by the injection of some contrast material: helical CT may very well allow us to image the whole vessel before the contrast from a single injection washes out and this may not be possible by the slower scanning modes. We point out that the CT scanner illustrated in Fig. 2 is in fact modern helical CT scanner.

For the sake of not over-complicating our discussion, in this chapter we restrict our attention (except where it is explicitly stated otherwise) to the problem of reconstructing two-dimensional objects from one-dimensional projections, rather than to what is done by modern helical cone-beam scanning (as in Fig. 8) and volumetric reconstruction. Schematically, the method of our data collection is shown in Fig. 9. The source and the detector strip are on either side of the object to be reconstructed and they move in unison around a common center of rotation denoted by O in Fig. 9. The data collection takes place in M distinct steps. The source and detector strip are rotated between two steps of the data collection by a small angle, but are assumed to be stationary while the measurement is taken. The M distinct positions of the source during the M steps of the data collection are indicated by the points $S_0, \ldots, S_{M-1}$ in Fig. 9. In simulating this geometry of data collection,

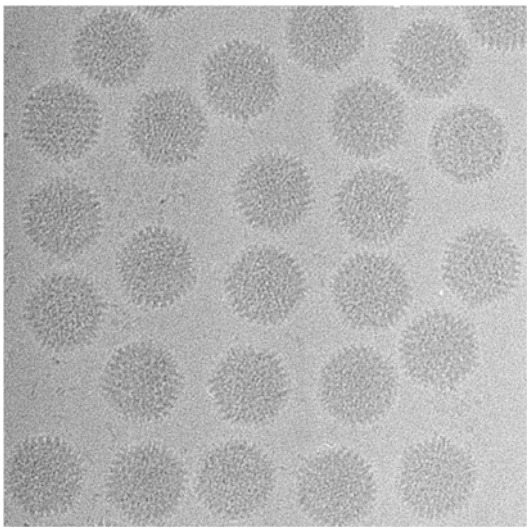


Fig. 6 Part of an electron micrograph containing projections of multiple copies of the human adenovirus type 5 (illustration provided by C. San Martín of the Centro Nacional de Biotecnología, Spain)

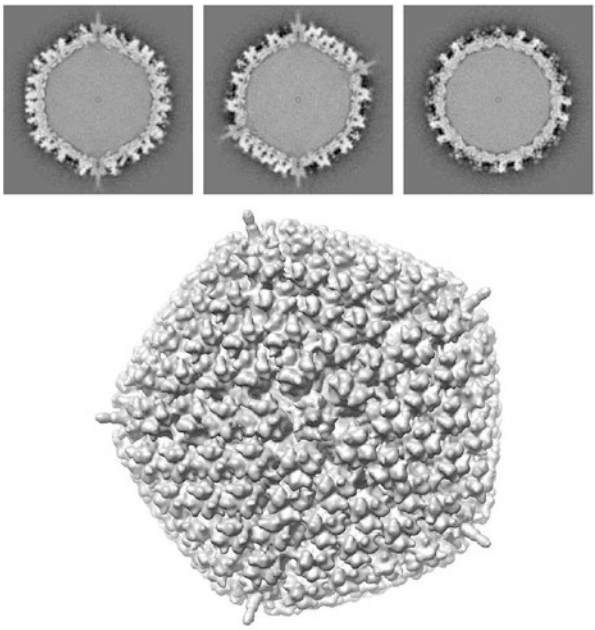


Fig. 7 *Top*: Reconstructed values, from electron microscopic data such as in Fig. 6, of the human adenovirus type 5 in three mutually orthogonal slices through the center of the reconstruction. *Bottom*: Computer graphic display of the surface of the virus based on the three-dimensional reconstruction (illustration provided by C. San Martín of the Centro Nacional de Biotecnología, Spain)

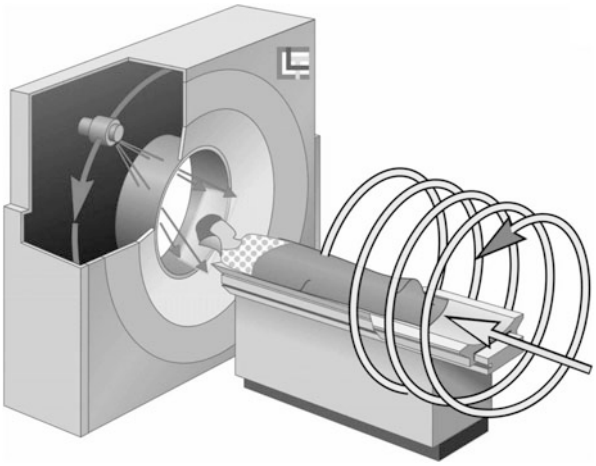


Fig. 8 Helical (also known as spiral) CT (illustration provided by G. Wang of the Virginia Polytechnic Institute & State University)

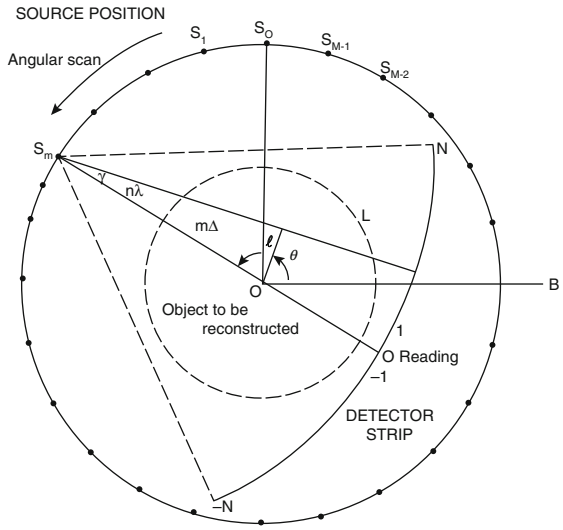


Fig. 9 Schematic of a standard method of data collection (divergent beam). This is consistent with the data collection mode for CT that is shown in Fig. 1 (reproduced from [23])

we assume that the source is a point source. The detector strip consists of $2N + 1$ detectors, spaced equally on an arc whose center is the source position. The line from the source to the center of rotation goes through the center of the central detector. (This description is that of the geometry that is assumed in much of what follows and it does not exactly match the data collection by any actual CT scanner. In particular, in real CT scanners the central ray usually does not go through the

middle of the central detector, as a 1/4 detector offset is quite common.) The object to be reconstructed is a picture such that its picture region (i.e., a region outside of which the values assigned to the picture are zero) is enclosed by the broken circle shown in Fig. 9. We assume that the origin of the coordinate system (with respect to which the picture values $\mu_{\tilde{e}}(x, y)$ are defined) is the center of rotation, O, of the apparatus.

Until now we have used $\mu_{\tilde{e}}(x, y)$ to denote the relative linear attenuation at the point (x, y), where (x, y) was in reference to a rectangular coordinate system, see Fig. 1. However, it is often more convenient to use polar coordinates. We use the phrase a *function of two polar variables* to describe a function f whose values $f(r, \phi)$ represent the value of some physical parameter (such as the relative linear attenuation) at the geometrical point whose polar coordinates are (r, ϕ).

We define the *Radon transform* $\mathscr{R}f$ of a function f of two polar variables as follows: for any real number pairs (ℓ, θ),

$$\begin{aligned} [\mathscr{R}f](\ell, \theta) &= \int_{-\infty}^{\infty} f\left(\sqrt{\ell^2 + z^2},\ \theta + \tan^{-1}(z/\ell)\right) dz, \quad \text{if } \ell \neq 0, \\ [\mathscr{R}f](0, \theta) &= \int_{-\infty}^{\infty} f(z, \theta + \pi/2)\, dz. \end{aligned} \tag{2}$$

Observing Fig. 1, we see that $[\mathscr{R}f](\ell, \theta)$ is the line integral of f along the line L. (Note that the dummy variable z in (2) does not exactly match the variable z as indicated in Fig. 1. In (2) $z = 0$ corresponds to the point where the perpendicular dropped on L from the origin meets L.)

In tomography we may assume that a *picture function* has bounded support; i.e., that there exists a real number E, such that $f(r, \phi) = 0$ if $r > E$. (E can be chosen as the radius of the broken circle in Fig. 9, which should enclose the square-shaped reconstruction region in Fig. 1.) For such a function, $[\mathscr{R}f](\ell, \theta) = 0$ if $\ell > E$.

The input data to a reconstruction algorithm are estimates (based on physical measurements) of the values of $[\mathscr{R}f](\ell, \theta)$ for a finite number of pairs (ℓ, θ); its output is an estimate, in some sense, of f. Suppose that estimates of $[\mathscr{R}f](\ell, \theta)$ are known for I pairs: $(\ell_1, \theta_1), \ldots, (\ell_I, \theta_I)$. For $1 \leq i \leq I$, we define $\mathscr{R}_i f$ by

$$\mathscr{R}_i f = [\mathscr{R}f](\ell_i, \theta_i). \tag{3}$$

In what follows we use y_i to denote the available estimate of $\mathscr{R}_i f$ and we use y to denote the I-dimensional vector whose ith component is y_i. We refer to the vector y as the *measurement vector*. When designing a reconstruction algorithm we assume that the method of data collection, and hence the set $\{(\ell_1, \theta_1), \ldots, (\ell_I, \theta_I)\}$, is fixed and known. The reconstruction problem is

given the data y, **estimate** the picture f.

We shall usually use f^* to denote the estimate of the picture f.

In the mathematical idealization of the reconstruction problem, what we are looking for is an operator $\mathscr{R}^{-1}$, which is an *inverse* of $\mathscr{R}$ in the sense that, for any picture function f, $\mathscr{R}^{-1}\mathscr{R}f$ is f (i.e., $\mathscr{R}^{-1}$ associates with the function $\mathscr{R}f$ the function f). Just as (2) describes how the value of $\mathscr{R}f$ is defined at any real number pair (ℓ, θ) based on the values f assumes at points in its domain, we need a formula that for functions p of two real variables defines $\mathscr{R}^{-1}p$ at points (r, ϕ). Such a formula is

$$\left[\mathscr{R}^{-1}p\right](r,\phi) = \frac{1}{2\pi^2}\int_0^{\pi}\int_{-E}^{E}\frac{1}{r\cos(\theta-\phi)-\ell}\,p_1(\ell,\theta)\,d\ell\,d\theta, \tag{4}$$

where $p_1(\ell, \theta)$ denotes the partial derivative of $p(\ell, \theta)$ with respect to ℓ; it is of interest to compare this formula with (1). That the $\mathscr{R}^{-1}$ defined in this fashion is indeed the inverse of $\mathscr{R}$ is proven, for example, in Section 15.3 of [23].

A major category of algorithms for image reconstruction calculate f^* based on (4), or on alternative expressions for the inverse Radon transform $\mathscr{R}^{-1}$. We refer to this category as *transform methods.* While (4) provides an exact mathematical inverse, in practice it needs to be evaluated based on finite and imperfect data using the not unlimited capabilities of computers. The essence of any transform method is a *numerical procedure* (i.e., one that can be implemented on a digital computer), which estimates the value of a double integral, such as the one that appears on the right-hand side of (4), from given values of $y_i = p(\ell_i, \theta_i)$, $1 \le i \le I$. A very widely used example of transform methods is the so-called *filtered backprojection* (FBP) algorithm. The reason for this name can be understood by looking at the right-hand side of (4): the inner integral is essentially a filtering of the projection data for a fixed θ and the outer integral backprojects the filtered data into the reconstruction region. However, the implementational details for the divergent beam data collection specified in Fig. 9 are less than obvious, the solution outlined below is based on [28].

The data collection geometry we deal with is also described in Fig. 10. The X-ray source is always on a circle of radius D around the origin. The detector strip is an arc centered at the source. Each line can be considered as one of a set of divergent lines (σ, β), where β determines the source position and σ determines which of the lines diverging from this source position we are considering. This is an alternative way of specifying lines to the (ℓ, θ) notation used previously (in particular in Fig. 1). Of course, each (σ, β) line is also an (ℓ, θ) line, for some values of ℓ and θ that depend on σ and β. We use $g(\sigma, \beta)$ to denote the line integral of f along the line (σ, β). Clearly,

$$g(\sigma,\beta) = [\mathscr{R}f](D\sin\sigma,\ \beta+\sigma). \tag{5}$$

As shown in Fig. 9, we assume that projections are taken for M equally spaced values of β with angular spacing Δ, and that for each view the projected values are sampled at $2N + 1$ equally spaced angles with angular spacing λ. Thus g is known at points $(n\lambda, m\Delta)$, $-N \le n \le N$, $0 \le m \le M-1$, and $M\Delta = 2\pi$.

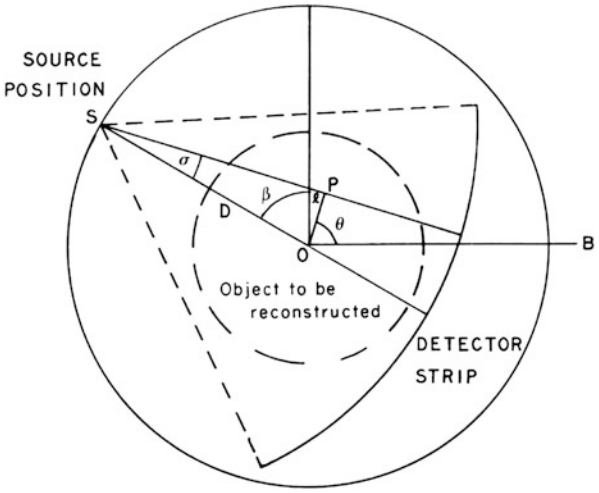


Fig. 10 Geometry of divergent beam data collection. Every one of the diverging lines is determined by two parameters β and σ. Let O be the origin and S be the position of the source, which always lies on a circle of radius D around O. Then $\beta + \pi/2$ is the angle the line OS makes with the baseline B and σ is the angle the divergent line makes with SO. The divergent line is also one of a set of parallel lines. As such it is determined by the parameters ℓ and θ. Let P be the point at which the divergent line meets the line through O that is perpendicular to it. Then ℓ is the distance from O to P and θ is the angle that OP makes with the baseline (reproduced from [22], Copyright 1981)

Even though the projection data consist of estimates (based on measurements) of $g(n\lambda, m\Delta)$, we use the same notation $g(n\lambda, m\Delta)$ for these estimates. The numerical implementation of the FBP method for divergent beams is carried out in two stages.

First we define, for $-N \leq n' \leq N$,

$$g_c(n'\lambda, m\Delta) = \lambda \sum_{n=-N}^{N} \cos(n\lambda) g(n\lambda, m\Delta) q^{(1)}\left((n'-n)\lambda\right) + \lambda \cos(n'\lambda) \sum_{n=-N}^{N} g(n\lambda, m\Delta) q^{(2)}\left((n'-n)\lambda\right). \quad (6)$$

The functions $q^{(1)}$ and $q^{(2)}$ determine the nature of the "filtering" in the filtered backprojection method. They are not arbitrary, but there are many possible choices for them, for a detailed discussion see Chapter 10 of [23]. Note that the first sum in (6) is a discrete convolution of $q^{(1)}$ and the projection data weighted by a cosine function, and the second sum is a discrete convolution of $q^{(2)}$ and the projection data.

Second we specify our reconstruction by

$$f^*(r,\phi) = \frac{D\Delta}{4\pi^2} \sum_{m=0}^{M-1} \frac{1}{W^2} g_c\left(\sigma', m\Delta\right), \tag{7}$$

where

$$\sigma' = \tan^{-1} \frac{r\cos(m\Delta - \phi)}{D + r\sin(m\Delta - \phi)}, \qquad -\frac{\pi}{2} \le \sigma' \le \frac{\pi}{2}, \tag{8}$$

and

$$W = \left((r\cos(m\Delta - \phi))^2 + (D + r\sin(m\Delta - \phi))^2\right)^{1/2}, \quad W > 0. \tag{9}$$

The meanings of σ' and W are that when the source is at angle $m\Delta$, the line that goes through (r,ϕ) is $(\sigma', m\Delta)$ and the distance between the source and $(r,\ \phi)$ is W. Implementation of (7) involves interpolation for approximating $g_c(\sigma', m\Delta)$ from values of $g_c(n'\lambda, m\Delta)$. The nature of such an interpolation is discussed in some detail in Section 8.5 of [23]. Note that (7) can be described as a "weighted backprojection." Given a point (r,ϕ) and a source position $m\Delta$, the line $(\sigma', m\Delta)$ is exactly the line from the source position $m\Delta$ through the point (r,ϕ). The contribution of the convolved ray sum $g_c(\sigma', m\Delta)$ to the value of f^* at points (r,ϕ) that the line goes through is inversely proportional to the square of the distance of the point (r,ϕ) from the source position $m\Delta$.

In this chapter we concentrate on the other major category of reconstruction algorithms, the so-called *series expansion methods.* In transform methods the techniques of mathematical analysis are used to find an inverse of the Radon transform. The inverse transform is described in terms of operators on functions defined over the whole continuum of real numbers. For implementation of the inverse Radon transform on a computer we have to replace these continuous operators by discrete ones that operate on functions whose values are known only for finitely many values of their arguments. This is done at the very end of the derivation of the reconstruction method. The series expansion approach is basically different. The problem itself is discretized at the very beginning: estimating the function is translated into finding a finite set of numbers. This is done as follows.

For any specified picture region, we fix a set of J *basis functions* $\{b_1, \ldots, b_J\}$. These ought to be chosen so that, for any picture f with the specified picture region that we may wish to reconstruct, there exists a linear combination of the basis functions that we consider an adequate approximation to f.

An example of such an approach is the $n \times n$ digitization in which we cover the picture region by an $n \times n$ array of identical small squares, called *pixels*. In this case $J = n^2$. We number the pixels from 1 to J, and define

$$b_j(r,\phi) = \begin{cases} 1, & \text{if } (r,\phi) \text{is inside the } j\text{th pixel,} \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

Then the $n \times n$ *digitization* of the picture f is the picture $\hat{f}$ defined by

$$\hat{f}(r,\phi) = \sum_{j=1}^{J} x_j b_j(r,\phi), \tag{11}$$

where x_j is the average value of f inside the jth pixel. A shorthand notation we use for equations of this type is $\hat{f} = \sum_{j=1}^{J} x_j b_j$.

Another (and usually preferable) way of choosing the basis functions is the following. *Generalized Kaiser–Bessel window functions*, which are also known by the simpler name *blobs,* form a large family of functions that can be defined in a Euclidean space of any dimension [40]. Here we restrict ourselves to a subfamily in the two-dimensional plane, whose elements have the form

$$b_{a,\alpha,\delta}(r,\phi) = \begin{cases} C_{a,\alpha,\delta}\left(1-\left(\frac{r}{a}\right)^2\right) I_2\left(\alpha\sqrt{1-\left(\frac{r}{a}\right)^2}\right), & \text{if } 0 \le r \le a, \\ 0, & \text{otherwise,} \end{cases} \tag{12}$$

where I_k denotes the modified Bessel function of the first kind of order k, a stands for the nonnegative radius of the blob, and α is a nonnegative real number that controls the shape of the blob. The multiplying constant $C_{a,\alpha,\delta}$ is defined below. Note that such a blob is circularly symmetric, since its value does not depend on ϕ. It has the value zero for all $r \ge a$ and its first derivatives are continuous everywhere. The "smoothness" of blobs can be controlled by the choice of the parameters a, α and δ, they can be made very smooth indeed as shown in Fig. 11.

For now let us consider the parameters a, α and δ, and hence the function $b_{a,\alpha,\delta}$, to be fixed. This fixed function gives rise to a set of J basis functions $\{b_1, \ldots, b_J\}$ as follows. We define a set $G = \{g_1, \ldots, g_J\}$ of *grid points* in the picture region. Then, for $1 \le j \le J$, b_j is obtained from $b_{a,\alpha,\delta}$ by shifting it in the plane so that its center is moved from the origin to g_j. This definition leaves a great deal of freedom in the selection of G, but it was found in practice advisable that it should consist of those points of a set (in rectangular coordinates)

$$G_\delta = \left\{ \left(\frac{m\delta}{2}, \frac{\sqrt{3}n\delta}{2} \right) \middle| \, m \text{ and } n \text{ are integers and } m+n \text{ is even} \right\} \tag{13}$$

that are also in the picture region. Here δ has to be a positive real number and G_δ is referred to as the *hexagonal grid with sampling distance* δ. Having fixed δ, we complete the definition in (12) by

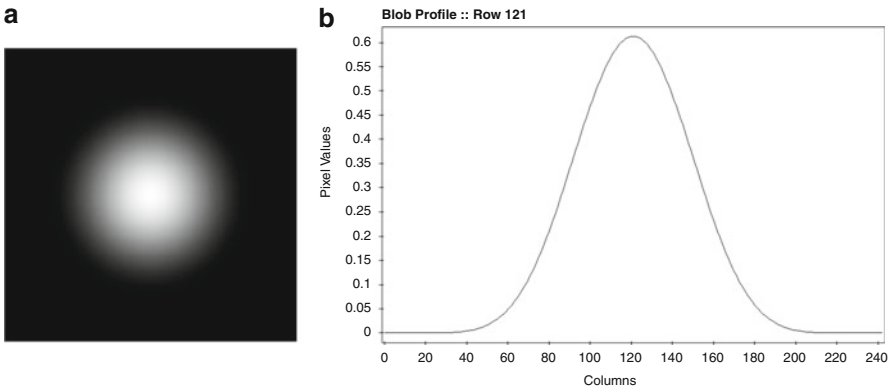


Fig. 11 (**a**) A 243 × 243 digitization of a blob. (**b**) Its values on the central row (reproduced from [23])

$$C_{a,\alpha,\delta} = \frac{\sqrt{3}\delta^2\alpha}{4\pi a^2 I_3(\alpha)}. \tag{14}$$

Pixel-based basis functions (10) have a unit value inside the pixels and zero outside. Blobs, on the other hand, have a bell-shaped profile that tapers smoothly in the radial direction from a high value at the center to the value 0 at the edge of their supports (i.e., at $r = a$ in (12)); see Fig. 11. The smoothness of blobs suggests that reconstructions of the form (11) are likely to be resistant to noise in the data. This has been shown to be particularly useful in fields in which the projection data are noisy, such as positron emission tomography and electron microscopy.

For blobs to achieve their full potential, the selection of the parameters a, α, and δ is important. When they are properly chosen [46], one can approximate homogeneous regions very well, in spite of the bell-shaped profile of the individual blobs. This is illustrated in Fig. 12b, in which a bone cross section shown in Fig. 12a is approximated by a linear combination of blob basis functions with the parameters $a = 0.1551, \alpha = 11.2829$, and $\delta = 0.0868$. There are some inaccuracies very near the sharp edges, but the interior of the bone is approximated with great accuracy. However, if we change the parameters ever so slightly to $a = 0.16, \alpha = 11.28$, and $\delta = 0.09$, then the best approximation that can be obtained by a linear combination of blob basis functions is shown in Fig. 12c, which is clearly inferior.

Irrespective of how the basis functions have been chosen, any picture $\hat{f}$ that can be represented as a linear combination of the basis functions b_j is uniquely determined by the choice of the coefficients x_j, $1 \le j \le J$, in the formula (11). We use x to denote the vector whose jth component is x_j and refer to x as the *image vector*.

It is easy to see that, under some mild mathematical assumptions,

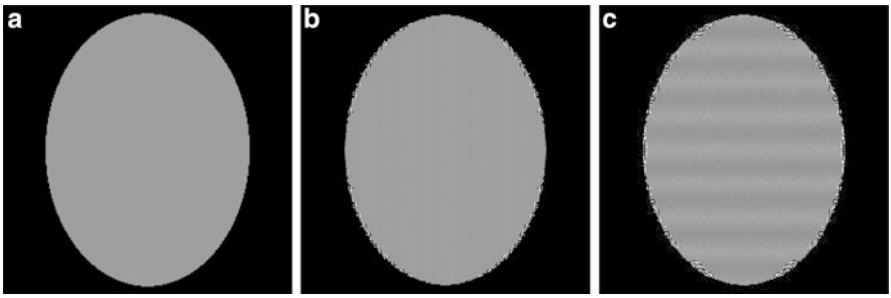


Fig. 12 (**a**) A 243 × 243 digitization of a bone cross section. (**b**) Its approximation with default blob parameters and (**c**) with slightly different parameters. The display window is very narrow for better indication of errors (reproduced from [23])

$$\mathscr{R}_i f \simeq \mathscr{R}_i \hat{f} = \sum_{j=1}^{J} x_j \mathscr{R}_i b_j, \tag{15}$$

for $1 \leq i \leq I$. Since the b_j are user-defined, usually the $\mathscr{R}_i b_j$ can be easily calculated by analytical means. For example, in the case when the b_j are defined by (10), $\mathscr{R}_i b_j$ is just the length of intersection with the jth pixel of the line of the ith position of the source-detector pair. We use $r_{i,j}$ to denote our calculated value of $\mathscr{R}_i b_j$. Hence,

$$r_{i,j} \simeq \mathscr{R}_i b_j. \tag{16}$$

Recall that y_i denotes the physically obtained estimate of $\mathscr{R}_i f$. Combining this with (15) and (16), we get that, for $1 \leq i \leq I$,

$$y_i \simeq \sum_{j=1}^{J} r_{i,j} x_j. \tag{17}$$

Let R denote the matrix whose (i, j)th element is $r_{i,j}$. We refer to this matrix as the *projection matrix*. Let e be the I-dimensional column vector whose ith component, e_i, is the difference between the left- and right-hand sides of (17). We refer to this as the *error vector*. Then (17) can be rewritten as

$$y = Rx + e. \tag{18}$$

The series expansion approach leads us to the following *discrete reconstruction problem*: based on (18),

given the data y, **estimate** the image vector x.

If the estimate that we find as our solution to the discrete reconstruction problem is the vector x^*, then the estimate f^* to the picture to be reconstructed is given by

$$f^* = \sum_{j=1}^{J} x_j^* b_j. \tag{19}$$

In (18), the vector e is unknown. The simple approach of trying to solve (18) by first assuming that e is the zero vector is dangerous: $y = Rx$ may have no solutions, or it may have many solutions, possibly none of which is any good for the practical problem at hand. Some criteria have to be developed, indicating which x ought to be chosen as a solution of (18). One way of doing this is by considering both the image vector x and the error vector e to be samples of random variables, denoted by X and E, respectively.

As an example of such an approach, let μ denote a J-dimensional vector of real numbers and let V denote a $J \times J$ positive definite symmetric matrix of real numbers. We can define a function p_X over the set of all J-dimensional vectors of real numbers by

$$p_X(x) = \frac{1}{(2\pi)^{J/2}(\det V)^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T V^{-1}(x-\mu)\right). \tag{20}$$

This p_X is a probability density function of a random variable X on the set of all J-dimensional vectors of real numbers whose mean vector is $\mu_x = \mu$ and whose covariance matrix is $V_X = V$. A random variable X defined in such a fashion is called a *multivariate Gaussian random variable*.

Let us now consider the random variables X and E associated with x and e of (18) without assuming any special form for them. In any case, p_X is referred to as the *prior probability density function*, since $p_X(x)$ indicates the likelihood of coming across an image vector similar to x. In CT it makes sense to adjust p_X to the area of the body we are imaging; the probabilities of the same picture representing a cross section of the head or of the thorax should be different. Based on p_X and p_E, a reasonable approach to solving the discrete reconstruction problem is: given the data y, choose the image vector x for which the value of

$$p_E(y - Rx)\, p_X(x) \tag{21}$$

is as large as possible. Note that the second term in the product is large for vectors x that have large prior probabilities, while the first term is large for vectors x that are consistent with the data (at least if p_E peaks at the zero vector). The relative importance of the two terms depends on the nature of p_X and p_E. If p_X is flat (many image vectors are equally likely) and p_E is highly peaked near the zero vector, then our criterion will produce an image vector x^* that fits the measured data y in the sense that Rx^* will be nearly the same as y. On the other hand, if p_E is flat (large errors are nearly as likely as small ones) but p_X is highly peaked, our having

made our measurements will have only a small effect on our preconceived idea as to how the image vector should be chosen. The x^* that maximizes (21) is called the *Bayesian estimate*. The discussion in this paragraph is quite general, since we have not assumed anything regarding the form of the random variables X and E.

If we assume that both X and E are multivariate Gaussian, then maximizing (21) becomes relatively simple. In that case it is easy to see from (20) that, assuming that μ_E is the zero vector, the x that maximizes (21) is the same x that minimizes

$$(y - Rx)^T V_E^{-1} (y - Rx) + (x - \mu_X)^T V_X^{-1} (x - \mu_X) . \tag{22}$$

When more precise information regarding the mean vector μ_X is not available, one can use for it a uniform picture, with an estimated (based on the projection data) average value assigned to every pixel; how well this works out in practice is illustrated below in the section on Numerical Methods and Case Examples. We also illustrate there an alternative choice that is appropriate for cardiac imaging in which μ_X is a time-averaged reconstruction. The noise model expressed by the first term of (22) is only approximate, but it is a reasonable accurate approximation of the effect of photon statistics in CT (see Section 3.1 of [23]).

As representative examples of the series expansion methods for image reconstruction we now discuss the*algebraic reconstruction techniques* (ART). All ART methods of image reconstruction are iterative procedures: they produce a sequence of vectors $x^{(0)}, x^{(1)}, \ldots$ that is supposed to *converge* to x^*. The process of producing $x^{(k+1)}$ from $x^{(k)}$ is referred to as an *iterative step*.

In ART, $x^{(k+1)}$ is obtained from $x^{(k)}$ by considering a single one of the I approximate equations, see (17). In fact, the equations are used in a *cyclic order*. We use i_k to denote $k (\text{mod } I) + 1$; i.e., $i_0 = 1$, $i_1 = 2, \ldots, i_{I-1} = I$, $i_I = 1$, $i_{I+1} = 2, \ldots$, and we use r_i to denote the J-dimensional column vector whose jth component is $r_{i,j}$. In other words, r_i is the transpose of the ith row of R. (In what follows we assume that, for $1 \leq i \leq I$, $\|r_i\|^2 = \langle r_i, r_i \rangle \neq 0$, where, as usual, $\|\bullet\|$ denotes the norm and $\langle \bullet, \bullet \rangle$ denotes the inner product.) An important point here is that this specification is incomplete because it depends on how we index the lines for which the integrals are estimated. As stated above, we assume that estimates of $[\mathscr{R} f](\ell, \theta)$ are known for I pairs: $(\ell_1, \theta_1), \ldots, (\ell_I, \theta_I)$. However, we have not specified the geometrical locations of the lines that are parametrized by these pairs. Since the order in which we do things in ART depends on the indexing i for the set of lines for which data are collected, the specification of ART as a reconstruction algorithm is complete only if it includes the indexing method for the lines, which we refer to as the *data access ordering*. We return to this point later on in this chapter.

A particularly simple variant of ART is the following.

$$\begin{aligned} &x^{(0)} \text{ is arbitrary,} \\ &x^{(k+1)} = x^{(k)} + c^{(k)} r_{i_k}, \end{aligned} \tag{23}$$

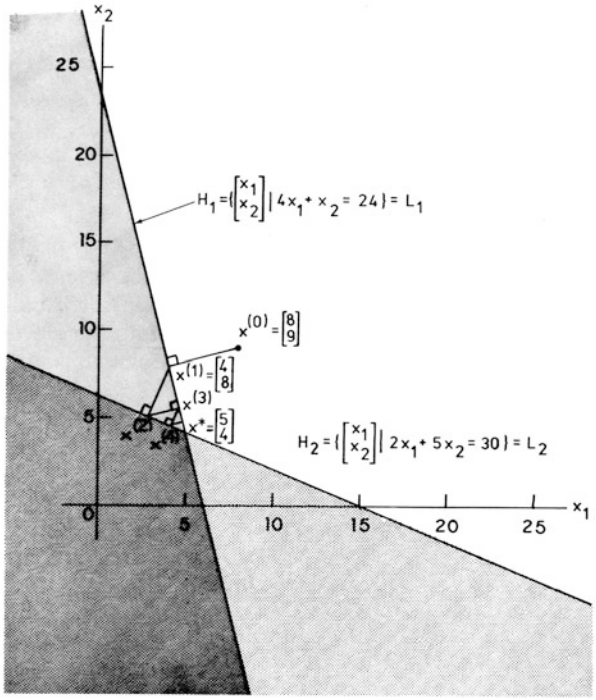


Fig. 13 Demonstration of the method of (23) and (24) (with $\lambda^{(k)} = 1$, for all k) for the simple case when $I = J = 2$ (illustration based on [25], Copyright 1976, with permission from Elsevier)

where

$$c^{(k)} = \lambda^{(k)} \frac{y_{i_k} - \left\langle r_{i_k} x^{(k)} \right\rangle}{\left\| r_{i_k} \right\|^2}, \tag{24}$$

with each $\lambda^{(k)}$ a real number, referred to as a *relaxation parameter*. It is easy to check that, for $k \geq 0$, if $\lambda^{(k)} = 1$, then

$$y_{i_k} = \sum_{j=1}^{J} r_{i_k, j} x_j^{(k+1)}, \tag{25}$$

i.e., the i_kth approximate equality is exactly satisfied after the kth step. This behavior is illustrated in Fig. 13 for a two-dimensional case with two equalities.

This method has an interesting, although by itself not particularly useful, mathematical property. Let

$$L = \{x | Rx = y\}. \tag{26}$$

A sequence $x^{(0)}, x^{(1)}, x^{(2)}, \ldots$ generated by (23) and (24) converges to a vector x^* in L, provided that L is not empty and that, for some ε_1 and ε_2 and for all k,

$$0 < \varepsilon_1 \leq \lambda^{(k)} \leq \varepsilon_2 < 2. \tag{27}$$

Furthermore, if x^0 is chosen to be the vector with zero components, then

$$\|x^*\| < \|x\|, \tag{28}$$

for all x in L other than x^*. A proof of this can be found in Section 11.2 of [23].

The reason why this result is not useful by itself is that the condition that L is not empty is unlikely to be satisfied in a real tomographic situation. However, as it is shown in Section 11.3 of [23], it can be used to derive an alternative ART algorithm that is useful in real applications, as we now explain.

Let us make the simplifying assumptions in (22) that V_X and V_E are both multiples of identity matrices of appropriate sizes. In other words, we assume that components of a sample of $X - \mu_X$ are uncorrelated, and that each component is a sample from the same Gaussian random variable; and we also assume that components of a sample of E are uncorrelated and that each component is a sample from the same zero mean Gaussian random variable. We use s^2 to denote the diagonal entries of V_X and n^2 to denote the diagonal entries of V_E and let $t = s/n$. According to (22), the Bayesian estimate is the vector x that minimizes

$$t^2 \|y - Rx\|^2 + \|x - \mu_X\|^2. \tag{29}$$

Note that a small value of t indicates that prior knowledge of the expected value of the image vector is important relative to the measured data, while a large value of t indicates the opposite. The following variant of ART converges to this Bayesian estimate, provided only that the condition expressed in (27) holds:

$$\begin{aligned} &u^{(0)} \text{ is the } I\text{-dimensional zero vector,} \\ &x^{(0)} = \mu_X, \\ &u^{(k+1)} = u^{(k)} + c^{(k)} e_{i_k}, \\ &x^{(k+1)} = x^{(k)} + t c^{(k)} r_{i_k}, \end{aligned} \tag{30}$$

where

$$c^{(k)} = \lambda^{(k)} \frac{t\left(y_{i_k} - \left\langle r_{i_k}, x^{(k)} \right\rangle\right) - u_{i_k}^{(k)}}{1 + t^2 \|r_{i_k}\|^2}. \tag{31}$$

Note that both in (23) and in (30) the updating of $x^{(k)}$ is very simple: we just add to $x^{(k)}$ a multiple of the vector r_{i_k}. In practice, this updating of $x^{(k)}$ can be computationally very inexpensive. Consider, for example, the basis functions associated with a digitization into pixels (10). Then $r_{i,j}$ is just the length of

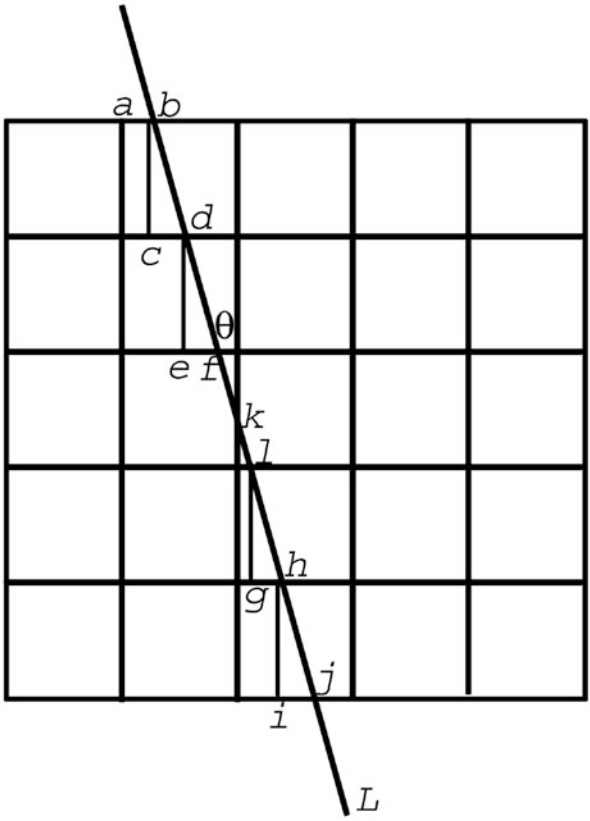


Fig. 14 A digital difference analyzer (DDA) for lines (reproduced from [23])

intersection of the ith line with the jth pixel. This has two consequences. First, most of the components of the vector r_{i_k} are zero. At most $2n-1$ pixels can be intersected by a straight line in an $n \times n$ digitization of a picture. Thus, of the n^2 components of r_{i_k}, at most $2n-1$ (and typically only about n) are nonzero. Second, the location and size of the nonzero components of r_{i_k} can be rapidly calculated from the geometrical location of the i_kth line relative to the $n \times n$ grid using a *digital difference analyzer* (DDA) methodology demonstrated in Fig. 14 (for details, see Section 4.6 of [23]). Thus, the projection matrix R does not need to be stored in the computer. Only one row of the matrix is needed at a time, and all information about this row is easily calculable. For this reason such methods are also referred to as *row-action methods.*

We investigate this point further, since it is basic to the understanding of the computational efficacy of ART. Suppose that we have obtained, using a DDA, the list $j_1, \ldots, j_U$ of indices such that $r_{i_k,j} = 0$ unless j is one of the $j_1, \ldots, j_U$. Then evaluation of $\langle r_{i_k}, x^{(k)} \rangle$ or of $\|r_{i_k}\|^2$ requires only U multiplications, which in our application is much smaller than J. The updating of x can be achieved by a further

U multiplications. This is because only those x_j need to be altered for which $j = j_u$ for some u, $1 \leq u \leq U$, and the alteration requires adding to $x_j^{(k)}$ a fixed multiple of $r_{i_k,j}$. This shows that a single step of either of the ART algorithms described above is very simple to implement in a computationally efficient way.

The ART are, in fact, a subclass of the class of *projection methods*, which have been demonstrated to be very effective in practice for solving convex feasibility problems with linear inequality constraints [6]. A recent advance in this direction is the development of the *superiorization methodology* [30] whose underlying idea is the following. There are many efficient iterative algorithms that produce feasible solutions for given constraints. Often the algorithm is perturbation resilient in the sense that, even if certain kinds of changes are made at the end of each iterative step, the algorithm still produces a feasible solution. This property is exploited in superiorization by using the perturbations to steer the algorithm to a solution that is not only constraints-compatible, but is also desirable according to an optimization criterion. This approach is applicable to many iterative procedures and optimization criteria.

4 Numerical Methods and Case Examples

Having seen that there is a variety of reconstruction algorithms, it is natural to ask for guidance as to when it is better to apply one rather than the others. Unfortunately, any general answer is likely to be misleading since the relative efficacy of algorithms depends on many things: the underlying task at hand, the method of data collection, the hardware/software available for implementing the algorithms, etc. The practical appropriateness of an algorithm under some specific circumstances needs experimental evaluation.

We are now going to illustrate this by comparing, from certain points of view, the various reconstruction algorithms mentioned in the previous section. Except where otherwise stated, the generation of images and their projection data, the reconstructions from such data, the evaluation of the results, and the graphical presentation of both the images and the evaluation results were done within the software package SNARK09 [12, 37].

We studied a cross section of a human head that was reconstructed by CT (see Fig. 15). Based on this cross section we described a skull enclosing the brain with ventricles, two tumors, and a hematoma (blood clot) using five ellipses, eight segments of circles, and two triangles. The tumors were placed so that they are vertically above the blood clot in the display. We used SNARK09 to obtain the density in each of 243×243 pixels of size 0.0752 cm. The resulting array of numbers is represented in Fig. 16. The nature of this display deserves careful discussion. The displayed values are linear attenuation coefficients $\mu_{\bar{e}}(x, y)$ at energy $\bar{e} = 60\,\text{keV}$ of the appropriate tissue types measured in cm^{-1}. Thus the values range between 0 (background, can be thought of as air) and 0.416 (bone of the skull). However, the interesting part of the picture is inside the skull. The values there range from

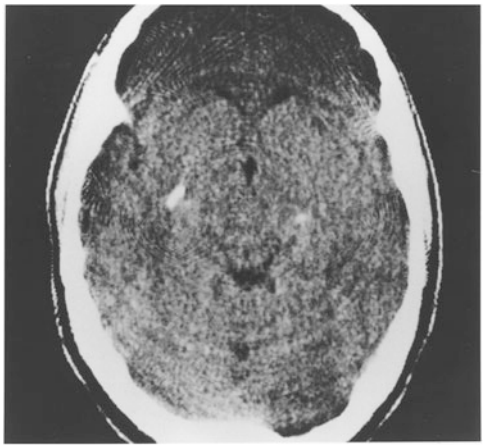


Fig. 15 Central part of an X-ray CT reconstruction of a cross section of the head of a patient. This served as the basis for our piecewise-homogeneous head phantom (reproduced from [23])

0.207 (cerebrospinal fluid) to 0.216 (metastatic breast tumor). The small differences between these tissues would not be noticeable if we used black to display zero, white to display 0.5, and corresponding grayness for values in between. To see clearly the features in the interior of the skull, we use zero (black) to represent the value 0.204 (or anything less) and 255 (white) to represent the value 0.21675 (or anything more). This way the small change in density by 0.001 corresponds to a change of 20 in display grayness, which is visible. We did this to produce Fig. 16 and the displays of all the reconstructions of the head phantoms used as illustrations in this chapter.

In Fig. 17a we show an actual brain cross section. The left half of the image shows a malignant tumor that has a highly textured appearance. In order to simulate the occurrence of a similarly textured object in our phantom we produced the phantom shown in Fig. 17b. Because of the medical relevance of imaging brains with such tumors, for the rest of this chapter we use the head phantom with this tumor added to it. (Due to our display method, it seems that there is a large range of values in the tumor. However, this is an illusion: the range of values in the tumor is less than 7 % of the range of values in the picture that is displayed in Fig. 16.)

One problem with the phantoms as defined so far is that a brain is far from being homogeneous: it has gray matter, white matter, blood vessels and capillaries carrying oxygenated blood to and deoxygenated blood from the brain, etc. This is even more so for bone, whose strength to a large extent is derived from its structural properties. There are methods that can obtain remarkably accurate reconstruction of piecewise homogeneous objects, but their performance may not be medically efficacious when applied to CT data from real objects with local inhomogeneities. So as not to fall into the trap of drawing too optimistic conclusions from experiments using piecewise homogeneous objects, we superimposed on our head phantom a random local variation that is obtained by picking, for each pixel, a sample

Fig. 16 A piecewise-homogeneous head phantom (reproduced from [23])

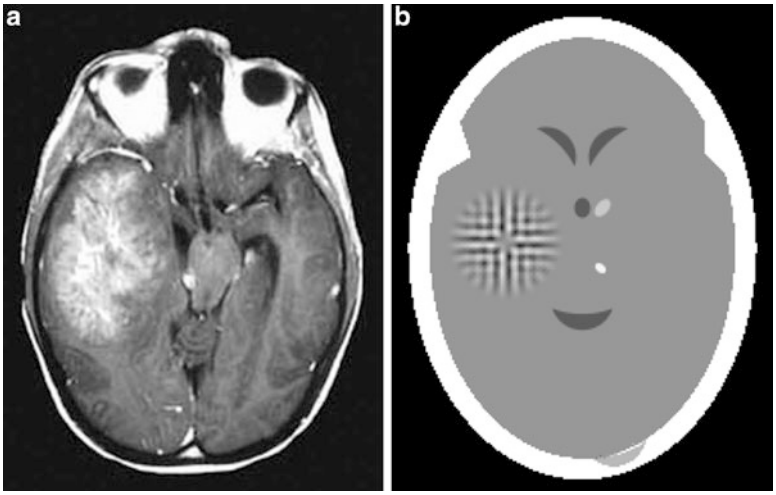


Fig. 17 (**a**) An actual brain cross section with a tumor (image is reproduced, with permission, from the Roswell Park Cancer Institute website). (**b**) The head phantom of Fig. 16 with a "large tumor" added to it (reproduced from [23])

from a Gaussian random variable X with mean $\mu_X = 1$ and standard deviation $\sigma_X = 0.0025$ and then multiplying the previously estimated linear attenuation coefficient at that energy level with that sample. In Fig. 18 we show the result of this.

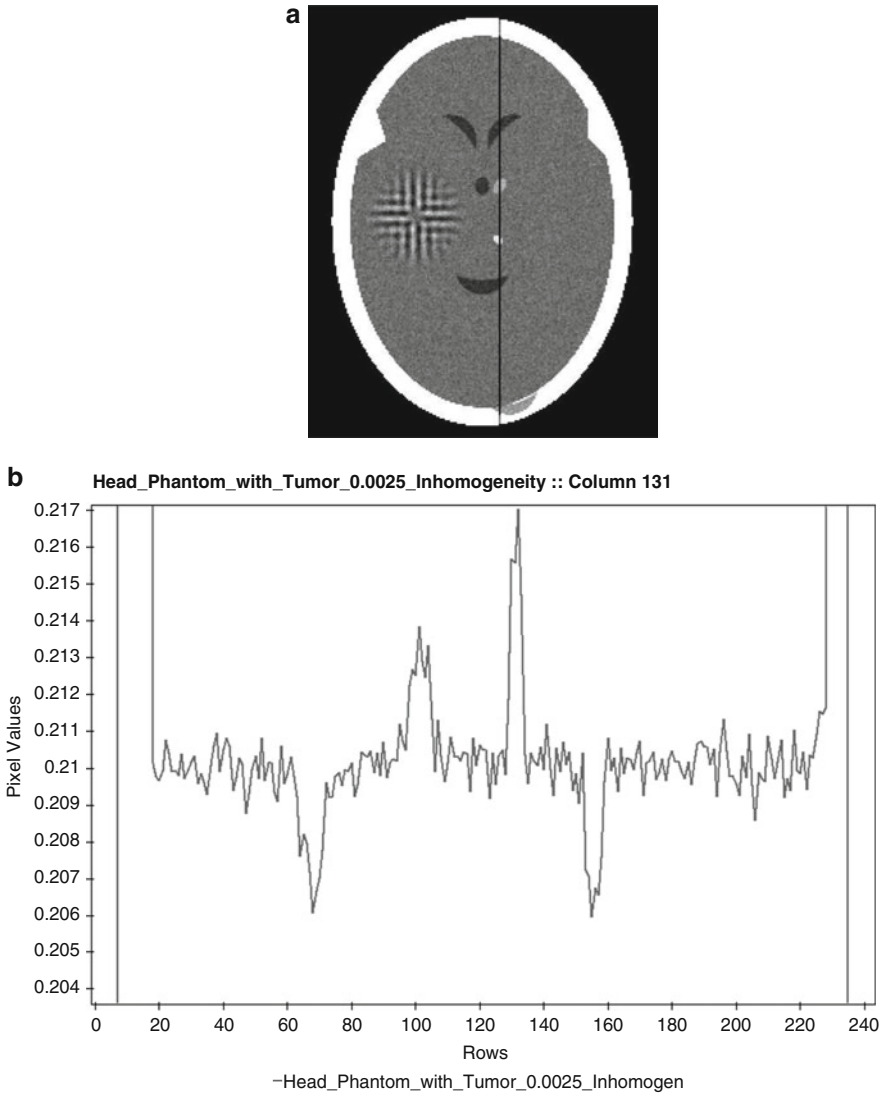


Fig. 18 (**a**) A head phantom with local inhomogeneities with the 131st of the 243 columns indicated by a *vertical line*. (**b**) The densities along this column in the phantom (reproduced from [23])

A reconstruction is a digitized picture. If it is a reconstruction from simulated projection data of a test phantom, we can judge its quality by comparing it with the digitization of the phantom. Naturally, both the picture region and the grid must be the same size for the reconstruction and the digitized phantom. We now discuss how to illustrate and measure the resemblance between a reconstruction and a phantom.

Visual evaluation is of course the most straightforward way. One may display both the phantom and the reconstruction and observe whether all features in which one is interested in the phantom are reproduced in the reconstruction and whether any spurious features have been introduced by the reconstruction process. A difficulty with such a qualitative evaluation is its subjectivity, people often disagree on which of two pictures resembles a third one more closely.

A more quantitative way of evaluating pictures is the following. Select a column of pixels that goes through a number of interesting features. For example, in our digitized head phantom the 131st of the 243 columns goes through the ventricles, both tumors, and the hematoma. In Fig. 18a we indicate this column. A way to evaluate the quality of a reconstruction is to compare the graphs of the 243 pixel densities for this column in the phantom (shown in Fig. 18b) and the reconstruction.

It also appears desirable to use a single value that provides a rough measure of the closeness of the reconstruction to the phantom. We now describe two different methods of doing this. In our definition of these two *picture distance measures* we use $t_{u,v}$ and $r_{u,v}$ to denote the densities of the vth pixel of the uth row of the digitized test phantom and the reconstruction, respectively, and $\bar{t}$ to denote the average of the densities in the digitized test phantom. We assume that both pictures are $n \times n$. Let

$$d = \left(\sum_{u=1}^{n} \sum_{v=1}^{n} (t_{u,v} - r_{u,v})^2 / \sum_{u=1}^{n} \sum_{v=1}^{n} (t_{u,v} - \bar{t})^2 \right)^{1/2}. \tag{32}$$

$$r = \sum_{u=1}^{n} \sum_{v=1}^{n} |t_{u,v} - r_{u,v}| / \sum_{u=1}^{n} \sum_{v=1}^{n} |t_{u,v}|. \tag{33}$$

($|x|$ denotes the absolute value of x.) These are often-used measures in the literature.

These measures emphasize different aspects of picture quality. The first one, d, is a *normalized root mean squared distance measure*. A large difference in a few places causes the value of d to be large. Note that the value of d is 1 if the reconstruction is a uniformly dense picture with the correct average density. The second one, r, is a *normalized mean absolute distance measure*. As opposed to d, it emphasizes the importance of a lot of small errors rather than of a few large errors. Note that the value of r is 1 if the reconstruction is a uniformly dense picture with zero density.

However, a collection of a few numbers cannot possibly take care of all the ways in which two pictures may differ from each other. Rank ordering reconstructions based on a few measures of closeness to the phantom can be misleading. We recommend instead a *statistical hypothesis testing* based methodology that allows us to evaluate the relative efficacy of reconstruction methods for a given task.

This evaluation methodology considers the following to be the relevant basic question: given a specific medical problem, what is the relative merit of two (or more) image reconstruction algorithms in presenting images that are helpful for solving the problem? (Compare this with the alternative essentially unanswerable question: which is the best reconstruction algorithm?) Ideally, the evaluation should

be based on the performance of human observers. However, that is costly and complex, since a number of observers have to be used, each has to read many images, conditions have to be carefully controlled, etc. Such reasons lead us to use *numerical observers* instead of humans. The evaluation methodology consists of four steps:

(i) Generation of random samples from a statistically described ensemble of images (phantoms) representative of the medical problem and computer simulation of the data collection by the device under investigation.
(ii) Reconstruction from the data so generated by each of the algorithms.
(iii) Assignment of a *figure of merit* (FOM) to each reconstruction. The FOM should measure the usefulness of the reconstruction for solving the medical problem.
(iv) Calculation of *statistical significance* (based on the FOMs of all the reconstructions) by which the null hypothesis that the reconstructions are equally helpful for solving the problem at hand can be rejected.

We now discuss details. For relevance to a particular medical task, the steps must be adjusted to that task. The task for which comparative evaluations of various pairs of reconstruction algorithms are reported below is that of detecting small low-contrast tumors in the brain based on reconstructions from CT data.

The ensemble of images generated for this task is based on the head phantom with a large tumor and local inhomogeneities. Note that this by itself provides us a statistical ensemble because the local inhomogeneities are introduced using a Gaussian random variable. However there is an additional (for the task more relevant) variability within the ensemble that is achieved as follows. We specify a large number of pairs of potential tumor sites, the locations of the sites in a pair are symmetrically placed in the left and right halves of the brain. In any sample from the ensemble, exactly one of each pair of the sites will actually have a tumor placed there, with equal probability for either site. The tumors are circular in shape of radius 0.1 cm and with linear attenuation as for the meningioma in the original phantom. In Fig. 19a we illustrate one sample from this ensemble. Once a sample has been picked, we generate projection data for it by simulating a CT scanner, with all its physical inaccuracies as compared to the idealized Radon transform. (Such inaccuracies include: the finite number of measurements, statistical noise due to the finite number of X-ray photons used during the measurements, the hardening of the polychromatic X-ray beam as it passes through the body, the width of the detector, and the scattering of X-ray photons.) Further variability is introduced at this stage, since the data are generated by simulating noise due to photon statistics. In Fig. 19b we show a reconstruction from one such projection data set. The tumors are hard to see in this reconstruction, but that is exactly the point: we are trying to evaluate which of two reconstruction algorithms provides images in which the tumors are easier to identify. If we make the task too easy (by having large and/or high-contrast tumors), then all reasonable reconstruction algorithms would perform perfectly from the point of view of the task. On the other hand, if the task is too difficult (very small and very low-contrast tumors), then correct detection would become essentially a

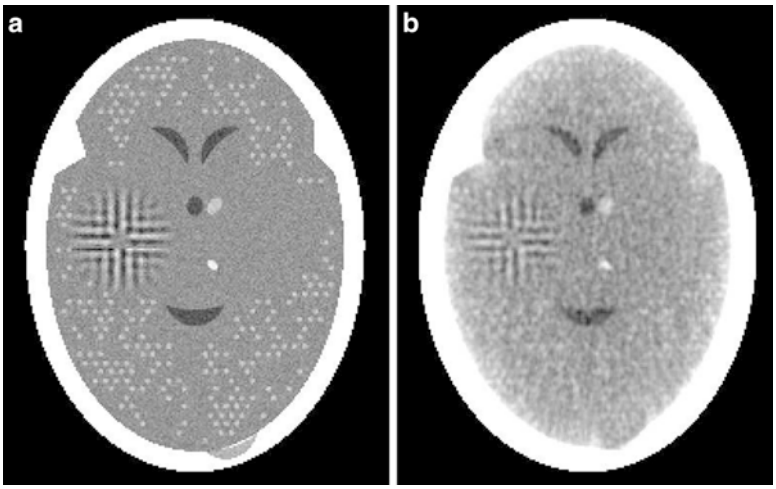


Fig. 19 (**a**) A random sample from the ensemble of phantoms for the task-oriented comparison of reconstruction algorithms. (**b**) A reconstruction from noisy projection data taken of the phantom illustrated in (**a**) (reproduced from [23])

matter of luck, rather than of algorithm performance. Our ensemble was chosen to be in-between these extremes. The FOM that we chose to use is specific to the type of ensemble of phantoms that we have just specified.

Given a phantom and one of its reconstructions, as in Fig. 19, we define the *image-wise region of interest FOM* (IROI) as

$$\mathrm{IROI} = \frac{\sum_{b=1}^{B} \left(\alpha_t^r(b) - \alpha_n^r(b)\right)}{\sqrt{\sum_{b=1}^{B} \left(\alpha_n^r(b) - \frac{1}{B}\sum_{b'=1}^{B} \alpha_n^r(b')\right)^2}} \Big/ \frac{\sum_{b=1}^{B} \left(\alpha_t^p(b) - \alpha_n^p(b)\right)}{\sqrt{\sum_{b=1}^{B} \left(\alpha_n^p(b) - \frac{1}{B}\sum_{b'=1}^{B} \alpha_n^p(b')\right)^2}}. \tag{34}$$

The specification of the terms in this formula is as follows. For any digitized picture and for any potential tumor site, let the *average density* in that picture for that site be the sum over all pixels whose center falls within the site of the pixel densities divided by the number of such pixels. Let us number the pairs of potential tumor sites from 1 to B, and let (for $1 \leq b \leq B$) $\alpha_t^p(b)$ (respectively, $\alpha_n^p(b)$) denote the average density in the phantom for site of the bth pair that has (respectively, has not) the tumor in it. We specify similarly $\alpha_t^r(b)$ (respectively, $\alpha_n^r(b)$), for the reconstruction. The first thing to note about the resulting formula (34) is that the numerator and the denominator in the big fraction are exactly the same except that the numerator refers to the reconstruction and the denominator refers to the phantom. Thus, if the reconstruction is perfect (in the sense of being identical to the phantom) then

IROI $= 1$. Analyzing the contents of the numerator and the denominator, we see that they are (except for constants that cancel out) the mean difference between the average values at the sites with tumors and the sites without tumors, divided by the standard deviation of the average values at the non-tumor sites. It has been found by experiments with human observers that this FOM correlates well with the performance of people [49].

In order to obtain statistically significant results we need to sample the ensemble of phantoms and generate projection data a number (say C) of times. (For the experiments reported below we used $C = 30$.) Suppose that we wish to compare the task-oriented performance of two reconstruction algorithms. For $1 \leq c \leq C$, let $\text{IROI}^1(c)$ and $\text{IROI}^2(c)$ denote the values of IROI, as defined by (34), for the reconstructions by the two algorithms from projection data of the cth phantom. The *null hypothesis* that the two reconstruction methods are equally good for the task at hand translates into the statistical statement that each value of $\text{IROI}^1(c) - \text{IROI}^2(c)$ is a sample of a continuous random variable D whose mean is 0. We have no idea of the shape of the probability density function p_D of this random variable, but by the central limit theorem (see, e.g., Section 1.2 of [23]), for a sufficiently large C,

$$s = \sum_{c=1}^{C} \left(\text{IROI}^1(c) - \text{IROI}^2(c)\right) \tag{35}$$

can be assumed to be a sample from a Gaussian random variable S with mean 0. This fact allows us to say (for details see Section 5.2 of [23]) that, at least approximately, S is a Gaussian random variable whose mean is 0 and whose variance is

$$V_S = \sum_{c=1}^{C} \left(\text{IROI}^1(c) - \text{IROI}^2(c)\right)^2. \tag{36}$$

It is a consequence of the null hypothesis that s is a sample from a zero-mean random variable. However, even if that were true, we would not expect our particular sample s to be exactly 0. Suppose for now that $s > 0$. This makes us suspect that in fact the first algorithm is better than the second one (for our task) and so the null hypothesis may be false. The question is: how significant is the observed value s for rejecting the null hypothesis? To answer this question we consider the *P-value*, which is the probability of a sample of S being as large or larger than s. If the null hypothesis were correct, we would not expect to come across an s defined by (35) for which the P-value is very small. Thus, the smallness of the P-value is a measure of *significance* for rejecting the null hypothesis that the two reconstruction algorithms are equally good for our task in favor of the *alternative hypothesis* that the first one is better than the second one. This is for the case when $s > 0$. If $s < 0$, then the P-value is the probability of a sample of S being as small or smaller than s and the alternative hypothesis is that the second algorithm is better than the first one.

Having specified various methodologies for reconstruction algorithm evaluation, we now apply them to specific algorithms. Whenever we report on the performance

of an algorithm for the reconstruction of a single two-dimensional phantom, the phantom is the one shown in Fig. 18. For experiments involving statistical hypothesis testing, we use the ensemble illustrated by Fig. 19. In either case, the data collection geometry is the one described in Fig. 9 with the number of source positions $M = 720$. Consequently, the angle $m\Delta$ shown in Fig. 9 is $0.5m$ degrees. The source positions are equally spaced around a circle of radius 78 cm. The distance of the source from the detector strip is 110.735 cm. There are 345 detectors, and the distance between two detectors along the arc of the detector strip is 0.10668 cm. We refer to this geometry of data collection as the *standard geometry.*

The reconstruction algorithm estimates a digitization of the phantom from the projection data. Figure 20 shows the 243×243 digitization of the head phantom, a reconstruction by FBP from perfect projections (line integrals) for the geometry just described, and the values of the digitized phantom and the reconstruction along the 131st column. The picture distance measures for this reconstruction are $d = 0.0531$ and $r = 0.0185$. Even though the data are perfect, the reconstruction is not. This is because a picture is not uniquely determined by its integrals along a finite number of lines. The best that a reconstruction algorithm can do is to *estimate* the picture.

There are interesting observations that one can make regarding this reconstruction. One is that, generally speaking, the brain appears smoother in it than in the phantom. This is because the FBP algorithm that we use was designed to perform efficaciously on real data and it does some smoothing to counteract the effect of noise. Consequently, small variations due to inhomogeneity are also smoothed. The most noticeable features in the reconstruction that are not present in the phantom are the streaks that seem to emanate from straight interfaces between the skull and the brain. (Similar features are observable in the real reconstruction shown in Fig. 15.) Their presence can be explained by considering Radon's formula (1), which expresses the distribution of the linear attenuation coefficient in terms of its line integrals. Consider an ℓ and a θ such that $m(\ell, \theta)$ is the integral along a line that is very near to a straight edge between the skull and the brain. Due to the fact that attenuation is much larger for bone than for brain, numerical estimation of the partial derivative $m_1(\ell, \theta)$ from the discretely sampled projection data is likely to be inaccurate, introducing errors into the calculated reconstruction. Phantoms that lack such anatomical features should not be used for algorithm evaluation, since the resulting reconstructions do not indicate the errors that will occur in a real application in which the object to be reconstructed is likely to have such straight interfaces. This is illustrated in Fig. 21.

The reconstructions shown in Figs. 20 and 21 are from "perfect" data; i.e., from line integrals based on the geometrical description of the phantoms. When data are collected by an actual CT scanner there are many physical reasons why the data so obtained can only provide approximations to such line integrals. In testing reconstruction algorithms we should use realistic projection data, which is what was done for the remaining two-dimensional reconstructions in this chapter. The exact method of simulated data collection (using SNARK09 [12, 37]) is described in Section 5.8 of [23], here we just give an outline. The data were collected for the head phantom shown in Fig. 20a according to the standard geometry. For

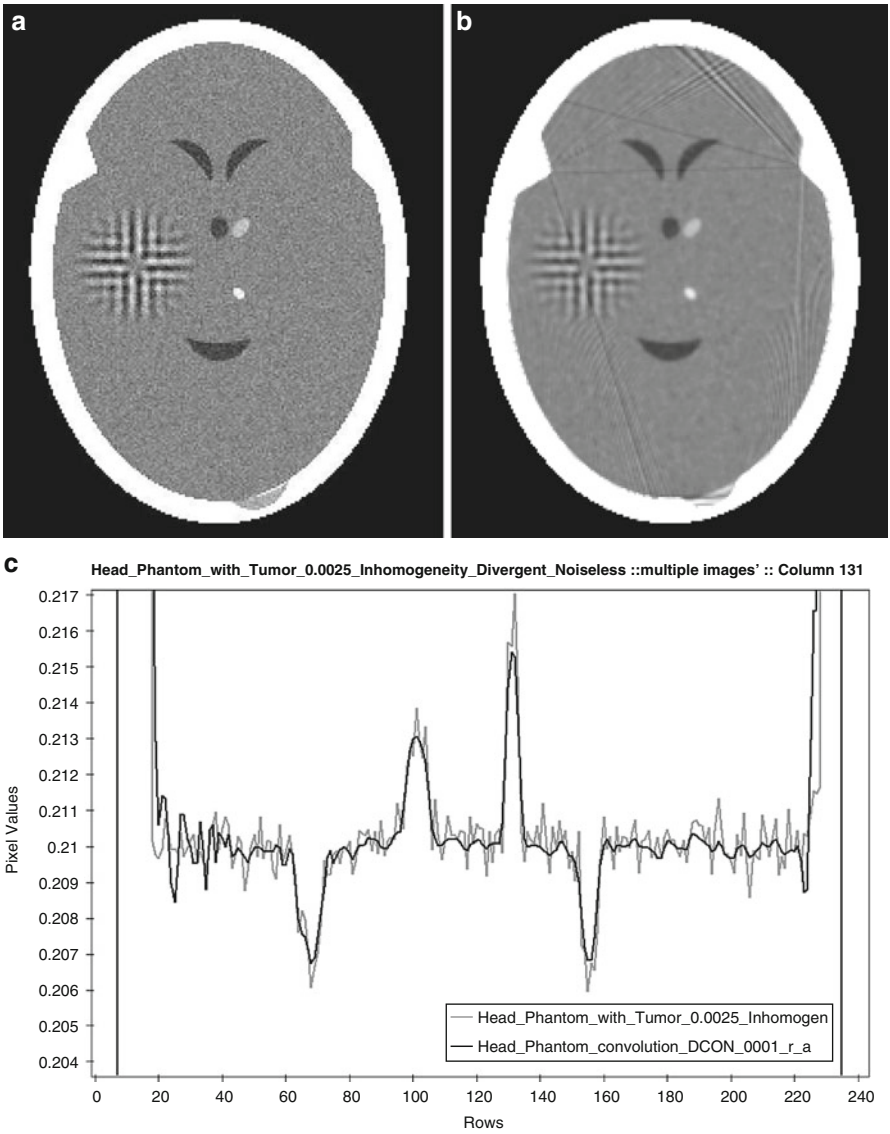


Fig. 20 (**a**) Head phantom (the same as Fig. 18a). (**b**) Its reconstruction from "perfect" data collected for the standard geometry. (**c**) Line plots of the 131st column of the phantom (*light*) and the reconstruction (*dark*) (reproduced from [23])

photon statistics we chose an average of million X-ray photons originating in the direction of each detector during the scanning of the head. A realistic spectrum of the polychromatic X-ray source was also simulated. The focal spot of the X-ray source was assumed to be a point, but the detectors were assumed to have width of

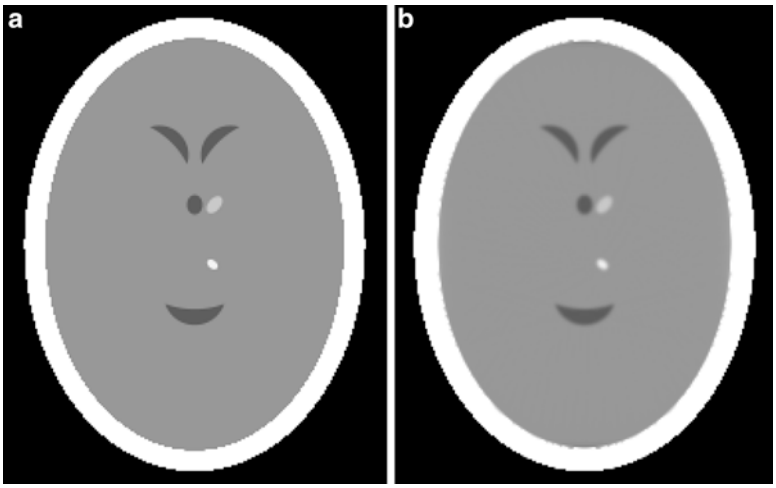


Fig. 21 (**a**) A simple head phantom without straight edges between bone and brain. (**b**) Its reconstruction from "perfect" data collected for the standard geometry. In this reconstruction there are no false features of the kind that emanate from the straight edges in Fig. 20b (reproduced from [23])

0.10668 cm (i.e., there are no gaps between the detectors). It was assumed that the number of scattered photons that are counted during the measurements is 5 % of the number of unscattered photons that are counted. The data so obtained was corrected for beam-hardening, to provide us with an estimate of the monochromatic projection data. The outcome of this correction is what we refer to as the *standard projection data*. For the experiments involving statistical evaluation, the same assumptions were made except that the phantom was randomly selected from the previously described ensemble; for an example, see Fig. 19a. Our illustrations are restricted to demonstrating the effects of various choices that can be made in ART and the comparison of ART with FBP.

We start with the variant of ART described by (23) and (24). We choose $x^{(0)}$ to represent a uniform picture, with the estimated (based on the standard projection data) average value of the phantom assigned to every pixel. (The estimation of the average value from projection data is described in Section 6.4 of [23].)

We first show that the order of equations in the system (the data access ordering discussed in the previous section) can have a significant effect on the practical performance of the algorithm, especially on the early iterates. With data collection such as the geometry depicted in Fig. 9, it is tempting to use the *sequential ordering*: access the data in the order $g(-N\lambda, 0), g((-N+1)\lambda, 0), \ldots, g(N\lambda, 0), g(-N\lambda, \Delta), g((-N+1)\lambda, \Delta), \ldots, g(N\lambda, \Delta), \ldots, \ldots, g(-N\lambda, (M-1)\Delta), g((-N+1)\lambda, (M-1)\Delta), \ldots, g(N\lambda, (M-1)\Delta)$, where $g(\sigma, \beta)$ denotes here the measured value of what is mathematically defined in (5). However, this sequential ordering is inferior to what is referred to as the *efficient ordering* in which

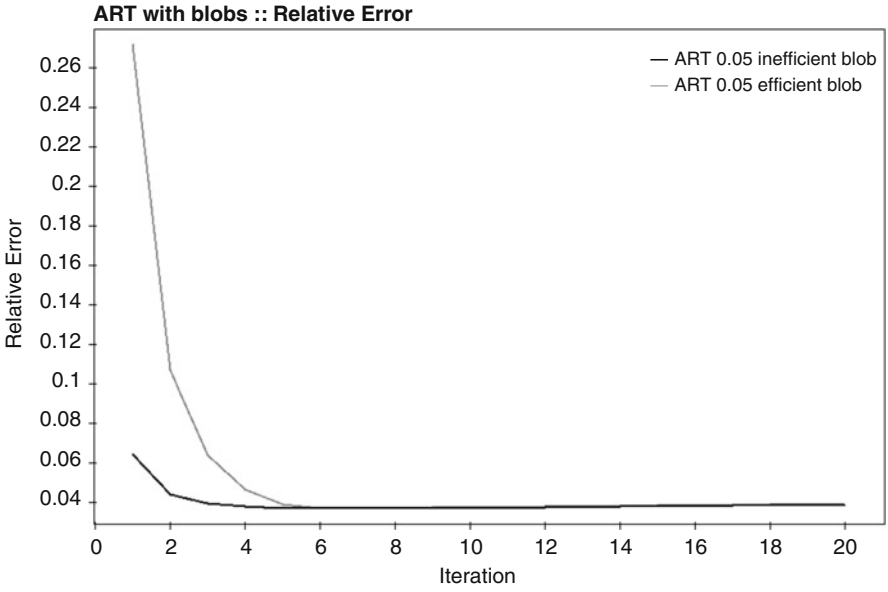


Fig. 22 Values of the picture distance measure r for ART reconstructions from the standard projection data with sequential ordering (*light*) and efficient ordering (*dark*), plotted at multiples of I iterations (reproduced from [23])

the order of projection directions $m\Delta$ and, for each view, the order of lines within the view is chosen so as to minimize the number of commonly intersected pixels by a line and the lines selected recently. This can be made mathematically precise by considering the decomposition into a product of prime numbers of M and of $2N+1$ [27]. SNARK09 [12,37] calculates the efficient order, but this is only useful if both M and of $2N+1$ decompose into several prime numbers, as is the case for our standard geometry for which $M = 720 = 2\times 2\times 2\times 2\times 3\times 3\times 5$ and $2N+1 = 345 = 3\times 5\times 23$. While the sequential ordering produces the sequences $m = 0, 1, 2, 3, 4, \ldots$ and $n = 0, 1, 2, 3, 4, \ldots$, the efficient ordering produces the sequences $m = 0, 360, 180, 540, 90, \ldots$ and $n = 0, 115, 230, 23, 138, \ldots$ These changes in data access ordering (keeping all other choices the same) translate into faster initial convergence of ART, as is illustrated in Fig. 22 by plotting the picture distance measure r of (33) against the number of times the algorithm cycled through all the data (all I equations). To produce this illustration we used blob basis functions and $\lambda^{(k)} = 0.05$, for all k. While it is clearly demonstrated that initially r gets reduced much faster with the efficient ordering, for the standard projection data it does not seem to matter much, since both orderings need about five cycles through the data to obtain a near-minimal value of r. In other applications in which the number of projection directions is much larger (e.g., in the order of 10,000 as is often the case in electron microscopy), one cycle through the data using the efficient ordering yields about as good a reconstruction as one is likely to get, but the

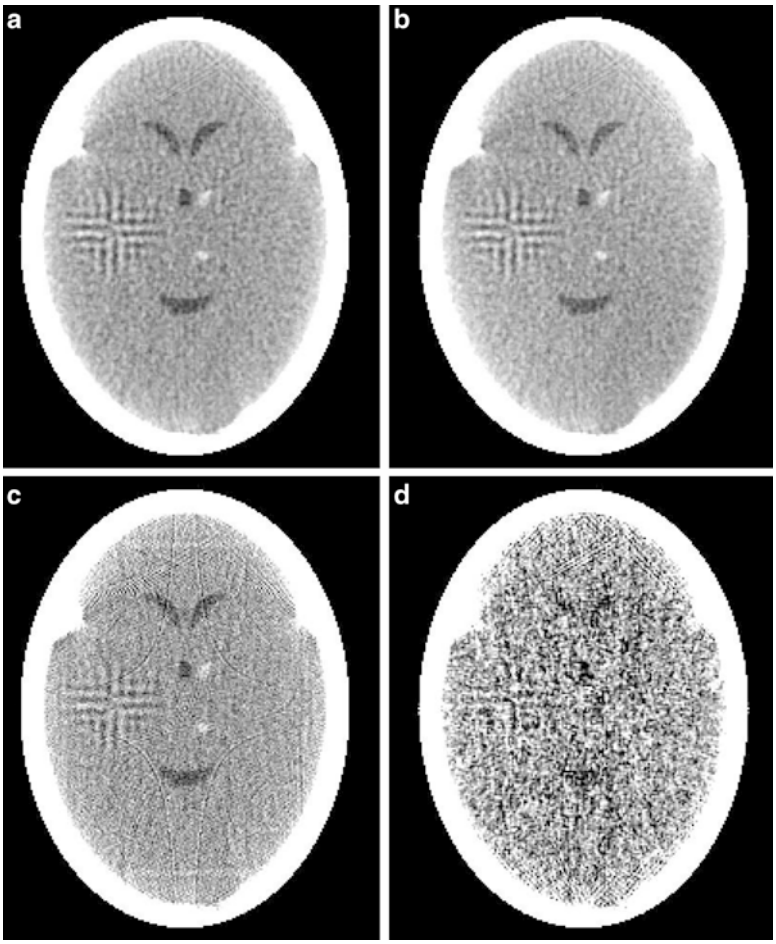


Fig. 23 Reconstructions from the standard projection data using ART. (**a**) ART with blobs, $\lambda^{(k)} = 0.05$, $5I$th iteration and efficient ordering. (**b**) ART with blobs, $\lambda^{(k)} = 0.05$, $5I$th iteration and sequential ordering. (**c**) ART with pixels, $\lambda^{(k)} = 0.05$, $5I$th iteration and efficient ordering. (**d**) ART with blobs, $\lambda^{(k)} = 1.0$, $2I$th iteration and efficient ordering (based on Fig. 11.4 of [23])

sequential ordering needs several cycles through the data. In addition, the efficacy of the reconstruction produced by the efficient ordering may very well be superior to that produced by the sequential ordering.

This is illustrated in Fig. 23 and Table 1. The reconstructions produced by the efficient and sequential orderings after five cycles through the data (the images of $x^{(5I)}$) are shown in Figs. 23a and b, respectively. Visually there is hardly any difference between them. This is confirmed by the picture distance measures in Table 1, they are only slightly better for the efficient ordering than for the sequential ordering. On the other hand, the execution time (within the SNARK09 [12, 37]

Table 1 Picture distance measures and timings (in seconds, of the implementations in SNARK09) for the reconstructions in Figs. 23 and 25

Reconstruction in	d	r	t	IROI
Figure 23a	0.0874	0.0373	163.7	0.1794
Figure 23b	0.0876	0.0391	148.9	0.1624
Figure 23c	0.0874	0.0470	29.2	0.1592
Figure 23d	0.0768	0.0488	66.2	0.1076
Figure 25b	0.1060	0.0423	8.7	0.1677

The last column reports the values, produced by a task-oriented evaluation experiment, of the IROI for the various algorithms (based on Table 11.1 of [23])

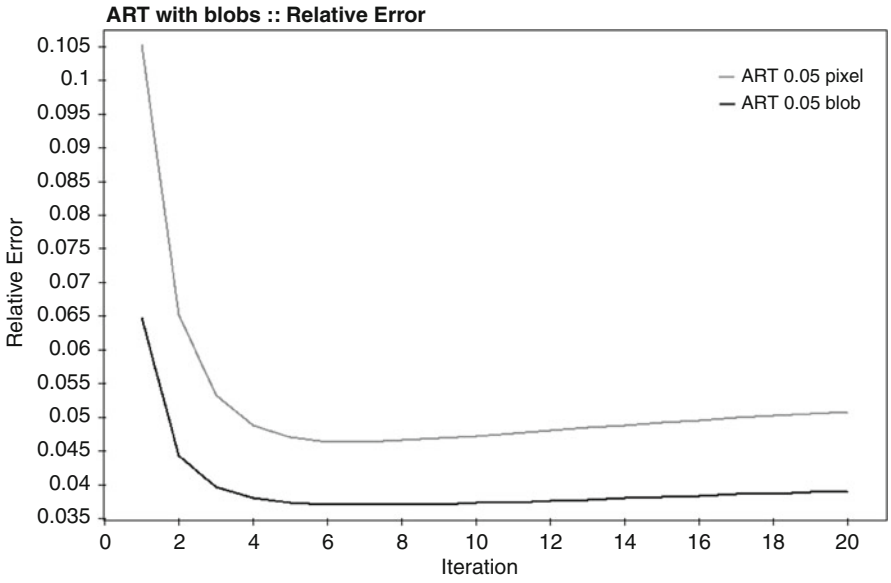


Fig. 24 Values of the picture distance measure r for ART reconstructions from the standard projection data with pixels (*light*) and blobs (*dark*), plotted at multiples of I iterations (reproduced from [23])

environment) is somewhat less for the sequential ordering. However, the task-oriented evaluation is unambiguous in its result: the IROI is larger for the efficient ordering and the associated P-value is less than 10^{-9}. This means that we can reject the null hypothesis that the two data access orderings are equally good in favor of the alternative hypothesis that the efficient ordering is better with extreme confidence.

Next we emphasize the importance of the basis functions. In Fig. 24 we plot the picture distance measure r against the number of times ART cycled through all the data, where we kept all other choices the same (in particular, efficient data access ordering and $\lambda^{(k)} = 0.05$, for all k). The two cases that we compare are when the basis functions are based on pixels (10) and when they are based on blobs (12). The results are impressive: as measured by r, blob basis functions are much better. The

result of the $5I$th iteration of the blob reconstruction is shown in Fig. 23a, while that of the $5I$th iteration of the pixel reconstruction is shown in Fig. 23c. The blob reconstruction appears to be clearly superior. In Table 1 we see a great improvement in the picture distance measure r but not in d. This reflects the fact, not visible in our display mode, that there are a few but relatively large errors in the blob reconstruction near the edges of the bone of the skull. From the points of view of the task-oriented figure of merit IROI, ART with blobs is found superior to ART with pixels with the relevant P-value less than 10^{-10}. As implemented in SNARK09, ART with blobs requires significantly more time than ART with pixels, but there exist more sophisticated implementations of ART with blobs that are much faster.

Underrelaxation is also a must when ART is applied to real, and hence imperfect, data. In the experiments reported so far $\lambda^{(k)}$ was set equal to 0.05 for all k. If we do not use underrelaxation (that is, we set $\lambda^{(k)}$ to 1 for all k), we get from the standard projection data the unacceptable reconstruction shown in Fig. 23d. Note that in this case we used the $2I$th iterate, further iterations give worse results. The reason for this is in the nature of ART: after one iterative step with $\lambda^{(k)} = 1$, the associated measurement is satisfied exactly as shown in (25) and so the process jumps around satisfying the noise in the measurements. Underrelaxation reduces the influence of the noise. The correct value of the relaxation parameter is application dependent; the noisier the data the more we should be underrelaxing. Note in Table 1 that the figure of merit IROI produced by the task-oriented study for the case without underrelaxation is much smaller than for the other cases.

Now we compare the best of our ART reconstruction (Fig. 23a, reproduced in Fig. 25a) with one produced by a carefully selected variant of FBP, see (6)–(9). For comparison, we show in Fig. 25b the reconstruction from our standard projection data obtained by FBP for divergent beams with linear interpolation and sinc window (also called the Shepp–Logan window, see [56]). For details of the meanings of these choices and the reasons for them, see Chapter 10 of [23]. The visual quality is similar to the best ART reconstruction. According to the picture distance measures in Table 1, ART is superior to FBP, and the same is true according to IROI with extreme significance (the P-value is less than 10^{-13}). This experiment confirms the reports in the literature that ART with blobs, underrelaxation and efficient ordering generally outperforms FBP in numerical evaluations of the quality of the reconstructions.

One thing though is indisputable: the ART with blob reconstruction took nearly 19 times longer than FBP. However, this should not be the determining factor, especially since the implementation of ART with blobs in SNARK09 is far from optimal and can be greatly improved. An advantage of ART over FBP is its flexibility. Even though until now we have reported its application only to data collected according to the standard geometry, ART is capable of reconstructing from data collected over any set of lines, as we soon demonstrate by an example of using ART for helical CT. FBP-type algorithms need to be reinvented for each new mode of data collection.

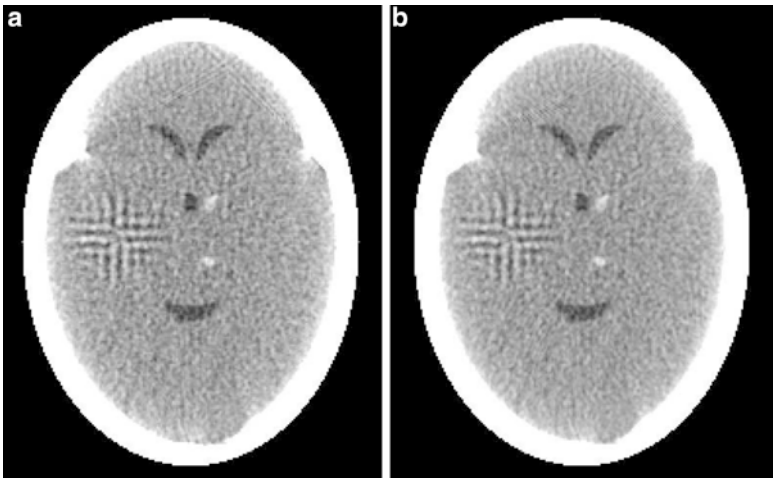


Fig. 25 Comparison of reconstructions from the standard projection data using (**a**) ART (the same as Fig. 23a) and (**b**) FBP (based on Fig. 11.4 of [23])

We now switch over to demonstrating the ART algorithm specified in (30) and (31). As stated before, that algorithm converges to the Bayesian estimate that is the minimizer of (29), provided that the condition expressed in (27) holds. This is the case if we set $\lambda^{(k)} = 0.05$, for all k, which is what we chose for the experiments on which we now report. The other choices that we made are blob basis functions, efficient ordering and that, in (29), $t = 10$ and μ_X represents a uniform picture with the estimated average value of the phantom assigned to every component.

There are alternative methods in the literature for minimizing (29), a particularly popular one is the method of *conjugate gradients* (CG); for a description of it that is appropriate for our context, see Section 12.5 of [23]. The CG method is also an iterative one, but one in which all the data are considered simultaneously in each iterative step. For this reason, the time of one iterative step of the CG method is approximately the same as that needed by ART for one cycle through all the data. In Fig. 26 we show a comparison of the picture distance measure r for CG and for ART.

Figure 26 and the picture distance measures in Table 2 imply that the quality of the reconstruction obtained by the 20th iterate of the conjugate gradient method should be as good as that obtained by the $5I$th iterate of additive ART. However this is not really so, as can be seen by looking at the reconstructed image in Fig. 27b. Indeed it needs another 20 iterations of the conjugate gradient method before the visual quality of the reconstruction matches that of the ART of (30) and (31) after $5I$ iterations, shown in Fig. 27a. So (for the standard projection data) the conjugate gradient method is not as fast as ART. This slower convergence of conjugate gradients relative to ART seems to be shared by other series expansion

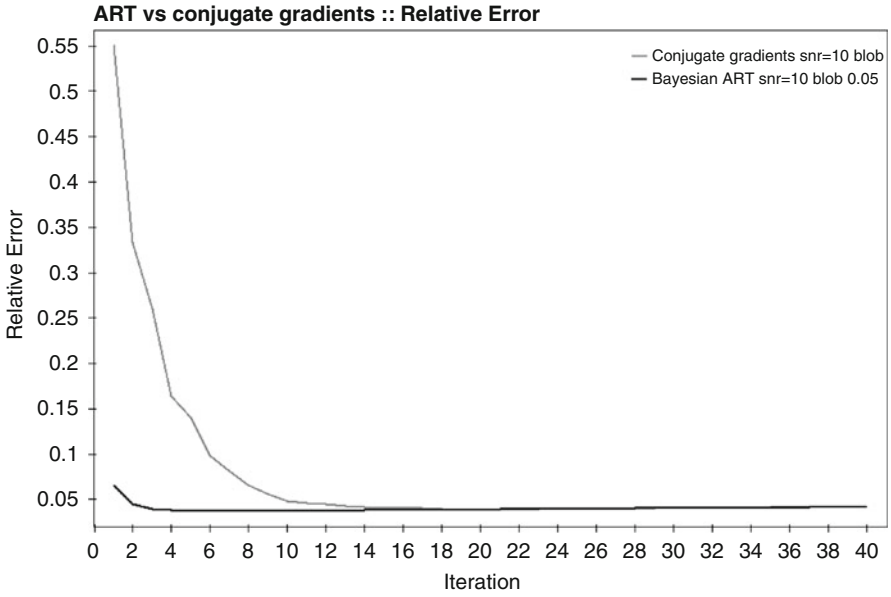


Fig. 26 Values of the picture distance measure r for reconstructions from the standard projection data using the conjugate gradient method (*light*) and ART (*dark*), plotted for comparable computational costs (reproduced from [23])

Table 2 Picture distance measures and timings (in seconds) for the reconstructions that minimize (29)

Algorithm	d	r	t
ART, $5I$th iterate	0.0878	0.0374	166.5
Conjugate gradient method, 20th iterate	0.0799	0.0387	489.1

Based on Table 12.2 of [23]

reconstruction methods that use all the data simultaneously in each iteration; see, for example, [58].

If we wish to reconstruct a three-dimensional body by the methods discussed till now, the only option available to us is to reconstruct the body cross section by cross section and then stack the cross sections to form the three-dimensional distribution. This may cause a number of problems, the most important of which are associated with time requirements. During the time needed to collect all the data, the patient may move, causing a misalignment between the cross sections. More basically, in moving organs such as the heart, changes in the organ over time are unavoidable, and it is usually not possible to collect data for all cross sections simultaneously.

Sometimes, it is actually the change in the object over time that is the desired information. If we wish to see cardiac wall motion, then it is essential that we reconstruct the whole three-dimensional object at short time intervals. One may consider this as a four-dimensional (spatio-temporal) reconstruction. One approach

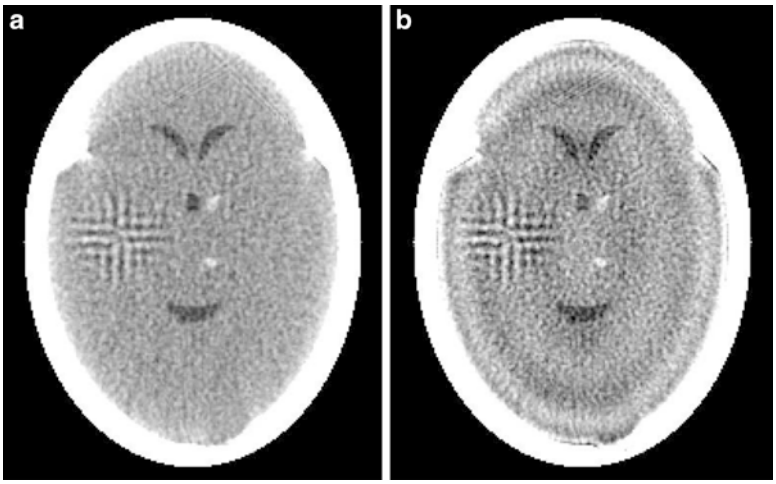


Fig. 27 Reconstructions from the standard projection data using iterative methods that minimize (29). (**a**) ART, 5*I* th iterate. (**b**) Conjugate gradient method, 20th iterate (based on Fig. 12.2 of [23])

to obtaining reconstructions of dynamically moving objects, such as the heart, from data that can be collected by helical CT (see Fig. 8) is to assume that the movement is cyclic. Assuming also that there exists a way of recording where we are in the cyclic movement as we take the 2D *views* of the moving 3D object, it is possible to bin the views into subsets such that all views that are binned into any one of the subsets have been taken at approximately the same phase of the cyclic movement, and so they are views of approximately the same (time frozen) 3D object. In the case of the heart this can be done by recording the electrocardiogram and noting on it the times when views have been taken. These views can then be binned, after the fact, according to the phases of the cardiac cycle.

We complete this section by giving a summary of such experiments, details can be found in Chapter 13 of [23]. The reconstructions were done by ART (here we made good use of the fact that ART does not require any particular arrangement of the lines for which the data were collected), using three-dimensional blobs [41] as the basis functions. We designed a phantom of the human thorax based on the description of the so-called FORBILD thorax phantom. We added to that stationary phantom two dynamically changing spheres representing the myocardium and a single contrast material filled cavity. We assumed that we are interested in this phantom at 24 equally spaced (in time) phases of the cardiac cycle. The first row of Fig. 28 shows a central cross section of this dynamic phantom at the two extremes of the 24 phases.

Projection taking was done by integrating the density of the phantom along lines between the X-ray source position and detectors in a two-dimensional array. For every source position, data were collected for 384 equally spaced detectors in each of 16 rows in the array. The size of each detector was assumed to be

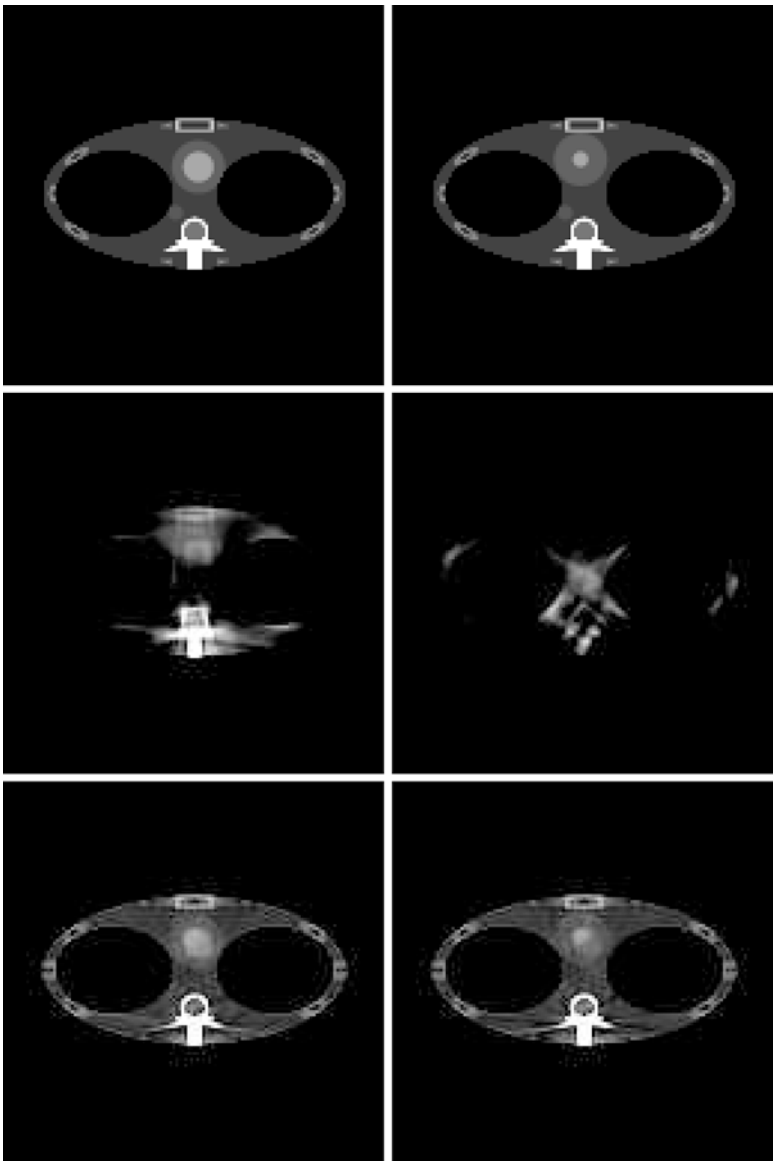


Fig. 28 The central cross section of the thorax phantom at the two extreme phases of the cardiac cycle. *First row*: the phantom. *Second row*: reconstruction from data collected at the time when the heart was in the appropriate phase after five cycles of simple ART. *Third row*: reconstruction from data collected at the time when the heart was in the appropriate phase after three cycles of Bayesian ART initialized with the reconstruction by two cycles of simple ART (based on Fig. 13.1 of [23])

0.425×0.425 cm. Data were collected (i.e., the pulsing of the X-ray source was simulated) at every 0.0015 second, using a total of 8,400 pulses. The number of turns of the helix in which the X-ray source moved during the data collection was 30. The radius of the helix was 57 cm, and the total movement parallel to the axis of the helix was 17.28 cm. The distance from the source to the detector array was 104 cm. Integrals of the density were collected for $I = 51{,}609{,}600$ rays (8,400 pulses times 16 rows of 384 detectors). Detector area and the effect of photon statistics were also simulated. The numbers used in this paragraph are not inappropriate for helical CT, but a state-of-the-art helical CT scanner would have more and smaller detectors and would be pulsed more frequently. In all our experiments we used $J = 2{,}153{,}935$ three-dimensional blobs to describe the reconstructed three-dimensional distributions.

In the first experiment we reconstructed the 24 phases of the cardiac cycle independently of each other. This was done by subdividing all the projection data into 24 subsets, each corresponding to one of the phases. A ray sum was put into a particular subset if it was collected due to a pulsing of the X-ray source at a time nearer to the central time for that phase than to the central time of any other phase. This results in a number of consecutive pulses producing data for the same phase and then there is a relatively large gap before the collected data are again used for that phase. This very nonuniform mode of data collection results in unacceptably bad reconstructions, two of which are demonstrated in the second row of Fig. 28. These reconstructions were produced using the simple ART of (23) and (24) with the three-dimensional blob basis functions, with all components of x^0 given the estimated average value based on the projection data, all $\lambda^{(k)} = 0.05$ and an efficient ordering. The results are shown at the end of the fifth cycle through the data associated with the particular phase of the cardiac cycle.

In the second experiment we used the other extreme: all the data were combined into a single projection data set, without any attention paid to the phases of the cardiac cycle. Because of the stationarity of most of the phantom and the overabundance of the projection data, we get (using the same choices for ART as in the previous paragraph) reconstructions that are good overall, but naturally the movement of the heart is blurred out due to the various views used in the reconstruction having been taken all through the cardiac cycle. We note that in this case there is no need to cycle through the data five times: the reconstruction at the end of the second cycle through the data is just about indistinguishable from the reconstruction at the end of the fifth cycle through the data.

However, our aim here is to see the dynamic changes in the heart. This can be achieved by using the Bayesian approach of (30) and (31). We selected in (29) μ_X as the reconstruction obtained at the end of the second ART cycle through all the data as described in the previous paragraph and $t = 0.8$. For each separate phase of the cardiac cycle, we used the algorithm specified by (30) and (31) for a further three cycles through the data that are associated with that particular phase. The relaxation parameter was again the constant 0.05. The results, for the two extreme phases of the cardiac cycle, are shown in the last row of Fig. 28. Here the overall reconstruction of the thorax is quite good and, at the same time, one can observe that the heart is

dynamically changing. With a state-of-the-art helical CT scanner (that would have more and smaller detectors and would be pulsed more often) we would get even better reconstructions.

5 Conclusion

Tomography is the process of producing an image of a distribution from estimates of its line integrals along a finite number of lines of known locations. There are a number of mathematical approaches to achieve this and we discussed and illustrated some of them. Of the investigated approaches, we found the performance of the method referred to as ART with blobs particularly good, especially if it is used with the appropriate data access ordering and relaxation parameters.

We subdivided the recommended readings into categories. For additional relevant information see [23] that has 280 references, 83 of which have been published since 2005.

Books related tomography [2, 7, 17–19, 23, 24, 29, 33, 50, 51, 59].

Papers on transform reconstruction methods and their applications [3, 9, 10, 13–15, 28, 34, 35, 52, 53, 56].

Papers on series expansion reconstruction methods and their applications [4, 5, 16, 20, 25, 27, 31, 32, 39–41, 44–46, 54, 57].

Papers on comparison of reconstruction methods [6, 12, 21, 30, 36, 37, 43, 47–49, 55, 58].

Papers on three-dimensional display of reconstructions [1, 8, 26, 42].

Cross-References

- Iterative Solution Methods
- Large-Scale Inverse Problems in Imaging
- Linear Inverse Problems
- Mathematical Methods in PET and SPECT Imaging
- Mathematics of Electron Tomography
- Regularization Methods for Ill-Posed Problems
- Mathematics of Photoacoustic and Thermoacoustic Tomography
- Variational Methods in Shape Analysis

References

1. Artzy, E., Frieder, G., Herman, G.T.: The theory, design, implementation and evaluation of a three-dimensional surface detection algorithm. Comput. Graph. Image Proc. **15**, 1–24 (1981)
2. Banhart, J.: Advanced Tomographic Methods in Materials Research and Engineering. Oxford University Press, Oxford (2008)

3. Bracewell, R.N.: Strip integration in radio astronomy. Aust. J. Phys. **9**, 198–217 (1956)
4. Browne, J.A., De Pierro, A.R.: A row-action alternative to the EM algorithm for maximizing likelihood in emission tomography. IEEE Trans. Med. Imaging **15**, 687–699 (1996)
5. Censor, Y., Altschuler, M.D., Powlis, W.D.: On the use of Cimmino's simultaneous projections method for computing a solution of the inverse problem in radiation therapy treatment planning. Inverse Prob. **4**, 607–623 (1988)
6. Censor, Y., Chen, W., Combettes, P.L., Davidi, R., Herman G.T.: On the effectiveness of projection methods for convex feasibility problems with linear inequality constraints. Comput. Optim. Appl. **51**, 1065–1088 (2012)
7. Censor, Y., Zenios, S.A.: Parallel Optimization: Theory, Algorithms and Applications. Oxford University Press, New York (1998)
8. Chen, L.S., Herman, G.T., Reynolds, R.A., Udupa, J.K.: Surface shading in the cuberille environment (erratum appeared in 6(2):67–69, 1986). IEEE Comput. Graph. Appl. **5**(12), 33–43 (1985)
9. Cormack, A.M.: Representation of a function by its line integrals, with some radiological applications. J. Appl. Phys. **34**, 2722–2727 (1963)
10. Crawford, C.R., King, K.F.: Computed-tomography scanning with simultaneous patient motion. Med. Phys. **17**, 967–982 (1990)
11. Crowther, R.A., DeRosier, D.J., Klug, A.: The reconstruction of a three-dimensional structure from projections and its application to electron microscopy. Proc. R. Soc. Lond. Ser.-A **A317**, 319–340 (1970)
12. Davidi, R., Herman, G.T., Klukowska, J.: SNARK09: A Programming System for the Reconstruction of 2D Images from 1D Projections (2009). http://www.dig.cs.gc.cuny.edu/software/snark09
13. DeRosier, D.J., Klug, A.: Reconstruction of three-dimensional structures from electron micrographs. Nature **217**, 130–134 (1968)
14. Edholm, P.R., Herman, G.T.: Linograms in image reconstruction from projections. IEEE Trans. Med. Imaging **6**, 301–307 (1987)
15. Edholm, P., Herman, G.T., Roberts, D.A.: Image reconstruction from linograms: implementation and evaluation. IEEE Trans. Med. Imaging **7**, 239–246 (1988)
16. Eggermont, P.P.B., Herman, G.T., Lent, A.: Iterative algorithms for large partitioned linear systems, with applications to image reconstruction. Linear Algebra Appl. **40**, 37–67 (1981)
17. Epstein, C.S.: Introduction to the Mathematics of Medical Imaging. 2nd edn. SIAM, Philadelphia (2007)
18. Frank, J.: Electron Tomography: Methods for Three-Dimensional Visualization of Structures in the Cell, 2nd edn. Springer, New York (2006)
19. Frank, J.: Three-Dimensional Electron Microscopy of Macromolecular Assemblies: Visualization of Biological Molecules in Their Native State. Oxford University Press, Oxford (2006)
20. Gordon, R., Bender, R., Herman, G.T.: Algebraic Reconstruction Techniques (ART) for three-dimensional electron microscopy and X-ray photography. J. Theory Biol. **29**, 471–481 (1970)
21. Hanson, K.M.: Method of evaluating image-recovery algorithms based on task performance. J. Opt. Soc. Am. A **7**, 1294–1304 (1990)
22. Herman, G.T.: Advanced principles of reconstructing algorithms. In: Newton, T.H., Potts, D.G. (eds.) Radiology of Skull and Brain. Technical Aspects of Computed Tomography, vol. 5, pp. 3888–3903. C.V. Mosby Company, St. Louis (1981)
23. Herman, G.T.: Fundamentals of Computerized Tomography: Image Reconstruction from Projections, 2nd edn. Springer, Berlin (2009)
24. Herman, G.T., Kuba, A.: Advances in Discrete Tomography and Its Applications. Birkhäuser, Boston (2007)
25. Herman, G.T., Lent, A.: Iterative reconstruction algorithms. Comput. Biol. Med. **6**, 273–294 (1976)
26. Herman, G.T., Liu, H.K.: Three-dimensional display of human organs from computed tomograms. Comput. Graph. Image Proc. **9**, 1–21 (1979)

27. Herman, G.T., Meyer, L.B.: Algebraic reconstruction techniques can be made computationally efficient. IEEE Trans. Med. Imaging **12**, 600–609 (1993)
28. Herman, G.T., Naparstek, A.: Fast image reconstruction based on a Radon inversion formula appropriate for rapidly collected data. SIAM J. Appl. Math. **33**, 511–533 (1977)
29. Herman, G.T., Tuy, H.K., Langenberg, K.J., Sabatier, P.C.: Basic Methods of Tomography and Inverse Problems. Institute of Physics Publishing, Bristol (1988)
30. Herman, G.T., Garduño, E., Davidi, R., Censor, Y.: Superiorization: an optimization heuristic for medical physics. Med. Phys. **39**, 5532–5546 (2012)
31. Hounsfield, G.N.: Computerized transverse axial scanning tomography: Part I, description of the system. Br. J. Radiol. **46**, 1016–1022 (1973)
32. Hudson, H.M., Larkin, R.S.: Accelerated image reconstruction using ordered subsets of projection data. IEEE Trans. Med. Imaging **13**, 601–609 (1994)
33. Kalender, W.A.: Computed Tomography: Fundamentals, System Technology, Image Quality, Applications, 2nd edn. Wiley-VCH (2006)
34. Kalender, W.A., Seissler, W., Klotz, E., Vock, P.: Spiral volumetric CT with single-breath-hold technique, continuous transport, and continuous scanner rotation. Radiology **176**, 181–183 (1990)
35. Katsevich, A.: Theoretically exact filtered backprojection-type inversion algorithm for spiral CT. SIAM J. Appl. Math. **62**, 2012–2026 (2002)
36. Kinahan, P.E., Matej, S., Karp, J.P., Herman, G.T., Lewitt, R.M.: Comparison of transform and iterative reconstruction techniques for a volume-imaging PET scanner with a large axial acceptance angle. IEEE Trans. Nucl. Sci. **42**, 2181–2287 (1995)
37. Klukowska, J., Davidi, R., Herman, G.T.: SNARK09 – a software package for the reconstruction of 2D images from 1D projections. Comput. Methods Programs Biomed. **110**, 424–440 (2013)
38. Lauterbur, P.C.: Medical imaging by nuclear magnetic resonance zeugmatography. IEEE Trans. Nucl. Sci. **26**, 2808–2811 (1979)
39. Levitan, E., Herman, G.T.: A maximum a posteriori probability expectation maximization algorithm for image reconstruction in emission tomography. IEEE Trans. Med. Imaging **6**, 185–192 (1987)
40. Lewitt, R.M.: Multidimensional digital image representation using generalized Kaiser–Bessel window functions. J. Opt. Soc. Am. A **7**, 1834–1846 (1990)
41. Lewitt, R.M.: Alternatives to voxels for image representation in iterative reconstruction algorithms. Phys. Med. Biol. **37**, 705–716 (1992)
42. Lorensen, W., Cline, H.: Marching cubes: a high-resolution 3D surface reconstruction algorithm. Comput. Graph. **21**(4), 163–169 (1987)
43. Maki, D.D., Birnbaum, B.A., Chakraborty, D.P., Jacobs, J.E., Carvalho, B.M., Herman, G.T.: Renal cyst pseudo-enhancement: beam hardening effects on CT numbers. Radiology **213**, 468–472 (1999)
44. Marabini, R., Rietzel, E., Schroeder, R., Herman, G.T., Carazo, J.M.: Three-dimensional reconstruction from reduced sets of very noisy images acquired following a single-axis tilt schema: application of a new three-dimensional reconstruction algorithm and objective comparison with weighted backprojection. J. Struct. Biol. **120**, 363–371 (1997)
45. Marabini, R., Herman, G.T., Carazo, J.-M.: 3D reconstruction in electron microscopy using ART with smooth spherically symmetric volume elements (blobs). Ultramicroscopy **72**, 53–65 (1998)
46. Matej, S., Lewitt, R.M.: Practical consideration for 3D image-reconstruction using spherically-symmetrical volume elements. IEEE Trans. Med. Imaging **15**, 68–78 (1996)
47. Matej, S., Herman, G.T., Narayan, T.K., Furuie, S.S., Lewitt, R.M., Kinahan, P.E.: Evaluation of task-oriented performance of several fully 3D PET reconstruction algorithms. Phys. Med. Biol. **39**, 355–367 (1994)
48. Matej, S., Furuie, S.S., Herman, G.T.: Relevance of statistically significant differences between reconstruction algorithms. IEEE Trans. Image Proc. **5**, 554–556 (1996)

49. Narayan, T.K., Herman, G.T.: Prediction of human observer performance by numerical observers: an experimental study. J. Opt. Soc. Am. A **16**, 679–693 (1999)
50. Natterer, F., Wübbeling, F.: Mathematical Methods in Image Reconstruction. SIAM, Philadelphia (2001)
51. Poulsen, H.F.: Three-Dimensional X-Ray Diffraction Microscopy: Mapping Polycrystals and Their Dynamics. Springer, Berlin (2004)
52. Radon, J.: Über die Bestimmung von Funktionen durch ihre Integralwerte längs gewisser Mannigfaltigkeiten. Ber. Verh. Sächs. Akad. Wiss., Leipzig, Math. Phys. Kl. **69**, 262–277 (1917)
53. Ramachandran, G.N., Lakshminarayanan, A.V.: Three-dimensional reconstruction from radiographs and electron micrographs: application of convolutions instead of Fourier transforms. Proc. Natl. Acad. Sci. USA **68**, 2236–2240 (1971)
54. Scheres, S.H.W., Gao, H., Valle, M., Herman, G.T., Eggermont, P.P.B., Frank, J., Carazo, J.-M.: Disentangling conformational states of macromolecules in 3D-EM through likelihood optimization. Nat. Methods **4**, 27–29 (2007)
55. Scheres, S.H.W., Nuñez-Ramirez, R., Sorzano, C.O.S., Carazo, J.M., Marabini, R.: Image processing for electron microscopy single-particle analysis using XMIPP. Nat. Protoc. **3**, 977–990 (2008)
56. Shepp, L.A., Logan, B.F.: The Fourier reconstruction of a head section. IEEE Trans. Nucl. Sci. **21**, 21–43 (1974)
57. Shepp, L.A., Vardi, Y.: Maximum likelihood reconstruction for emission tomography. IEEE Trans. Med. Imaging **1**, 113–122 (1982)
58. Sorzano, C.O.S., Marabini, R., Boisset, N., Rietzel, E., Schröder, R., Herman, G.T., Carazo, J.M.: The effect of overabundant projection directions on 3D reconstruction algorithms. J. Struct. Biol. **133**, 108–118 (2001)
59. Udupa, J.K., Herman, G.T.: 3D Imaging in Medicine, 2nd edn. CRC Press, Boca Raton (1999)

Microlocal Analysis in Tomography

Venkateswaran P. Krishnan and Eric Todd Quinto

Contents

1 Introduction 848
2 Motivation 848
X-Ray Tomography (CT) and Limited Data Problems 849
Electron Microscope Tomography (ET) Over Arbitrary Curves 853
Synthetic-Aperture Radar Imaging 856
General Observations 859
3 Properties of Tomographic Transforms 859
Function Spaces 859
Basic Properties of the Radon Line Transform 860
Continuity Results for the X-Ray Transform 863
Filtered Backprojection (FBP) for the X-Ray Transform 864
Limited Data Algorithms 866
ROI Tomography 866
Limited Angle CT 867
Fan Beam and Cone Beam CT 867
Algorithms in Conical Tilt ET 868
4 Microlocal Analysis 871
Singular Support and Wavefront Set 871
Pseudodifferential Operators 875
Fourier Integral Operators 880
5 Applications to Tomography 884
Microlocal Analysis in X-Ray CT 884
Limited Data X-Ray CT 887
Exterior X-Ray CT Data 887

V.P. Krishan
Centre for Applicable Mathematics, Tata Institute for Fundamental Research, Bangalore, Karnataka, India
e-mail: venkyp.krishnan@gmail.com; vkrishnan@math.tifrbng.res.in

E.T. Quinto (✉)
Department of Mathematics, Tufts University, Medford, MA, USA
e-mail: todd.quinto@tufts.edu

© Springer Science+Business Media New York 2015

O. Scherzer (ed.), *Handbook of Mathematical Methods in Imaging*,
DOI 10.1007/978-1-4939-0790-8_36

Limited Angle Data 888
Region of Interest (ROI) Data 888
Microlocal Analysis of Conical Tilt Electron Microscope Tomography (ET) 889
SAR Imaging 891
6 Conclusion 895
Cross-References 896
References 897

Abstract

Several limited data problems in tomography will be presented in this chapter, including ones for X-ray tomography, electron microscopy, and radar imaging. First, reconstructions from limited data will be evaluated to observe their strengths and weaknesses. Then, the basic analytic properties of the transforms will be presented. The concept of microlocal analysis will be introduced to make the notion of singularity precise. Finally, the microlocal properties of the tomographic transforms are given and then used to explain the observed strengths and limitations of the reconstructions. This will show that these limitations are intrinsic to these limited data problems themselves.

1 Introduction

In this chapter, a range of tomography problems are introduced, including X-ray imaging, limited data problems, electron microscopy, and radar imaging. The goal is to recover the singular features of the medium or object rather than to develop exact inversion formulas. Toward this end, microlocal analysis is used to understand the strengths and limitations inherent in these tomography problems. Microlocal analysis aids researchers in understanding those singular features that can be stably recovered, which could be very important when only limited or partial data are available. Furthermore, it helps explain the presence of artifacts in certain image reconstruction methods. In some cases, it might help to distinguish the true singularities from the false ones. These are the themes emphasized in this chapter.

In Sect. 2, the tomography problems are introduced. Reconstructions are presented for each problem and reconstruction quality is evaluated with the goal of finding strengths and limitations for each method. In Sect. 3, some basic properties of tomographic transforms are introduced. Microlocal analysis is introduced in Sect. 4. Finally, several applications in tomography and radar imaging are given in Sect. 5 with the goal of emphasizing the microlocal properties of these transforms. This powerful tool is used to clarify the strengths and limitations that are really intrinsic to the data.

2 Motivation

Several modalities in tomography, including X-ray tomography, limited data tomography, electron microscope tomography, and synthetic-aperture radar (SAR), are introduced in this section. For each problem, some history is given and then

reconstructions using such data are evaluated for strengths and weaknesses. The goal of this section is to observe, for each problem, object features that are well reconstructed and features that are not. These reconstructions are provided to motivate the study of microlocal analysis, which will be used in Sect. 5 to explain these reconstructions.

X-Ray Tomography (CT) and Limited Data Problems

In the 1970s, X-ray tomography revolutionized diagnostic medicine. For the first time, doctors were able to get clear and accurate pictures of the inside of the body without doing exploratory surgery. One part of this story began in the early 1960s. At that time, Allan Cormack consulted as a medical physicist at the Groote Schuur hospital in Cape Town, South Africa, and he checked whether X-ray machines were calibrated properly. He felt that there should be more information in the X-ray data than just what is obtained from single pictures, which project all organs onto the same plane, and he believed that X-rays could be used more effectively. He posited that if one takes X-ray images from multiple directions, one should be able to piece together the internal structure of the body. He then developed two algorithms [10, 11] for the problem. To give a proof of concept, he built a prototype scanner that showed his second algorithm was effective. Along with Godfrey Hounsfield of EMI in England, he received the 1979 Nobel Prize in Medicine. You can read more about him in the excellent biography [97].

X-ray CT is now used routinely in medicine and in industrial nondestructive testing, and it allows doctors to image the internal structure of the body without exploratory surgery. The basic physics and mathematical model are now described. Let ℓ be a line along which X-rays travel, and for $x \in \ell$, let $I(x)$ be the intensity (number of photons) at the point x. Let $f(x)$ be the attenuation coefficient of the body at x. For monochromatic light, f is proportional to the density at x, and by using a scale factor, they become the same. Beer's law [66] states that the decrease in intensity at x is proportional to the intensity $I(x)$, and the proportionality constant is $-f(x)$:

$$\frac{dI}{dx} = -f(x)I(x). \tag{1}$$

This makes sense heuristically because the more dense the material at x (i.e., the larger $f(x)$ is), the more the beam is attenuated and the greater the decrease of I at x. Equation (1) is a simple differential equation for I that can be solved using separation of variables. If I_0 is the intensity at the X-ray emitter – the point $x_0 \in \ell$ – and I_1 is the intensity at the detector, $x_1 \in \ell$, then one can integrate (1) to find

$$\ln\left(\frac{I_0}{I_1}\right) = \int_{x_0}^{x_1} f(x)\,\mathrm{d}x = \int_{x\in\ell} f(x)\,\mathrm{d}x\,.$$

This leads to the definition

$$\mathcal{R}_L(f)(\ell) = \int_{x\in\ell} f(x)\,\mathrm{d}x$$

where in this case, $\mathrm{d}x$ is the arc length measure on ℓ. The transform $\mathcal{R}_L$ was studied by the Austrian mathematician Johann Radon [84] in the early twentieth century because it was intriguing pure mathematics. This transform is called the *Radon line transform* (or X-ray transform).

To proceed mathematically, more notation is given. Let $\omega \in S^1$ and let $p \in \mathbb{R}$. Then, the line

$$\ell(\omega, p) = \{x \in \mathbb{R}^2 : x \cdot \omega = p\} \tag{2}$$

is perpendicular to ω and contains $p\omega$. Sometimes it will be useful to let ω be a function of polar angle $\varphi \in \mathbb{R}$,

$$\omega(\varphi) = (\cos(\varphi), \sin(\varphi)) \ .$$

In this parameterization

$$\mathcal{R}_L f(\omega, p) = \int_{x \in \ell(\omega,p)} f(x)\, \mathrm{d}x = \int_{t \in \mathbb{R}} f(p\omega + t\omega^{\perp})\, \mathrm{d}t \tag{3}$$

where $\omega^{\perp}$ is the unit vector $\pi/2$ radians counterclockwise from ω. This integral is defined for $f \in C_c(\mathbb{R}^2)$, and in fact $\mathcal{R}_L$ is continuous in a number of norms (see section "Continuity Results for the X-Ray Transform"). The basic properties of this transform are proven in Sect. 3.

First, consider the forward problem and a simple case that will show in a naive sense how the X-ray transform detects object boundaries.

Example 1. Let f be the characteristic function of the unit disk in $\mathbb{R}^2$. Then, using the Pythagorean Theorem, one sees that

$$\mathcal{R}_L f(\omega, p) = \begin{cases} 2\sqrt{1-p^2} & |p| \le 1 \\ 0 & |p| > 1 \end{cases} . \tag{4}$$

The function $\mathcal{R}_L f(\omega, p)$ in (4) is smooth except at $p = \pm 1$, that is, except for lines $\ell(\omega, \pm 1)$ as can be seen from Fig. 1. The data are not smooth at those lines and those lines are tangent to the boundary of the disk. This suggests that lines tangent to boundaries give special information about the specimen. In Sect. 4, the reader will discover what is mathematically special about those lines, and this will be related back to limited data tomography in Sect. 5.

For complete data, that is, data over all lines through the object, good reconstruction methods such as filtered backprojection (Theorem 9) are effective to reconstruct from X-ray CT data.

However, one cannot obtain complete data in many important tomography problems. These are called *limited data tomography problems*, and several important ones are now described. The goal at this point is to observe how the reconstructions look compared to the original objects.

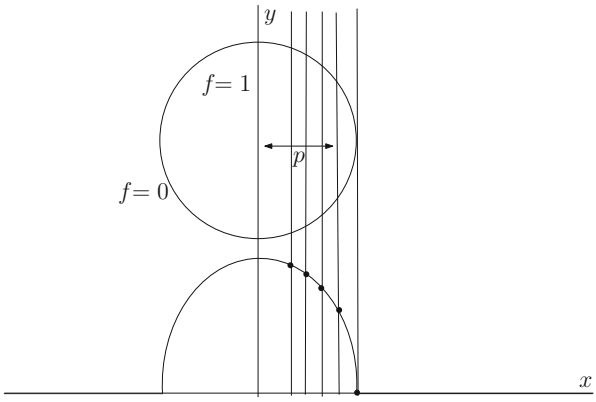


Fig. 1 This graph shows the calculation of the Radon transform in (4). The unit disk is above the graph. For $|p| \leq 1$, the Pythagorean theorem shows that the length of the intersection of $\ell(\omega, p)$ and the disk is $2\sqrt{1-p^2}$

Here are some guidelines as you read this section. For each problem and reconstruction, conjecture what is special about the object boundaries that are well reconstructed (and those that are badly reconstructed) in relation to the limited data set used.

Exterior X-Ray CT Data

Exterior CT data are data for lines that are outside an excluded region. Typically, that region is a circle of radius $r > 0$, so lines $\ell(\omega, p)$ for $|p| \geq r$ are in the data set. Theorem 5 in the next section shows that compactly supported functions can be uniquely reconstructed outside the excluded region from exterior data.

The exterior problem came about in the early days of tomography for CT scans around the beating heart. In those days, a single scan of a planar cross section could take several minutes, and movement of the heart would create artifacts in the scan. If an excluded region were chosen to contain the heart and be large enough so the outside of that region would not move, then data exterior to that region would be usable. However, scanners soon began to use fan beam data (see section "Fan Beam and Cone Beam CT"), and data could be acquired much more quickly. If the data acquisition is timed (gated), then data are acquired while the heart is in the same position over several heartbeats. Because more data can be taken more quickly with fan beam data, the heart can now be imaged using newer scanners, and movement of the heart is not as large a problem.

Exterior data are still important for imaging large objects such as rocket shells. Even with an industrial CT scanner, the X-rays will not penetrate the thick center of the rocket [88]. However, they can penetrate the outer rocket shell, and this gives exterior data.

One can recover functions of compact support from exterior data, at least outside the excluded region (see Theorem 5). Effective inversion methods were developed for exterior data by researchers including Bates and Lewitt [3], Natterer [65], Quinto [77, 79], and a stability analysis using singular value decompositions was done in [58].

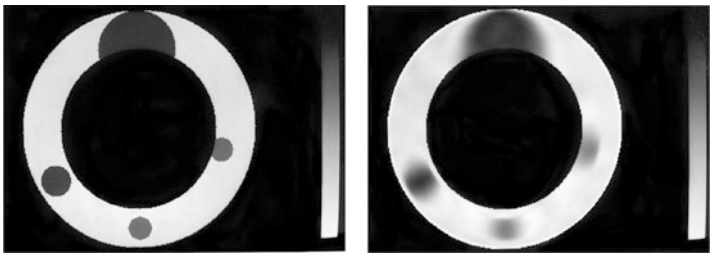


Fig. 2 Exterior reconstruction. Phantom (*left*) and reconstruction (*right*) from simulated data [77, ©IOP Publishing. Reproduced by permission of IOP Publishing. All rights reserved]. The outer diameter of the annulus is 1.5 times the inner diameter

Figure 2 is a reconstruction from exterior data: integrals are given over lines that do not meet the black central disk. The reconstruction method uses a singular value decomposition for the exterior Radon transform that includes a null space; it recovers the component of the object in the orthogonal complement of the null space and does an extrapolation to recover the null space component [77].

Note how some boundaries of the small circles are clearly reconstructed and others are not. In this case, how can you describe the boundaries that are well reconstructed in relation to the data set? Another question is whether the fuzzy boundaries are fuzzy because the algorithm is bad or could there be an additional explanation.

Allan Cormack's first algorithm [10] solved the exterior problem, but the algorithm was not numerically effective. The integrals in his algorithm were difficult to evaluate numerically with any accuracy because the integrand grew too rapidly. Other mathematicians tried to improve this method, but it was difficult. Because of this problem, Cormack developed a second method that uses full data and that gave good reconstructions [11].

It would be useful to know if limitations of Quinto's and Cormack's algorithms are problems with their algorithms or reflect something intrinsic to this limited data problem.

Limited Angle Data

Limited angle tomography is a classical problem from the early days of tomography [3, 60, 61]. In this case, data are given over all lines in a limited range of directions, or data for $\{(\omega(\varphi), p) : \varphi \in (-\Phi, \Phi), p \in \mathbb{R}\}$ where $\Phi \in (0, \pi/2)$. One can uniquely recover compactly supported functions from limited angle data, but this is not true for arbitrary functions (see Theorem 3).

Limited angle tomography is used in certain luggage scanners in which the X-ray source is on one side of the luggage and the detectors are on the other, and they move in opposite directions. Limited angle data are used in important current problems including dental X-ray scanning [68] and tomosynthesis (a tomographic technique to image breasts using transmitter and receiver that move on opposite

sides of the breast) [72]. Other algorithms were developed for this problem such as [12, 26, 49, 68].

The reconstruction in Fig. 3 is from limited angle data. Data are taken over all lines $\ell(\omega(\varphi), p)$ for $p \in \mathbb{R}$ and $\varphi \in [-\pi/4, \pi/4]$.

The algorithm used in this reconstruction is a truncated filtered backprojection (FBP) algorithm which is given in (26). Some boundaries in this reconstruction are well reconstructed and others are not. How do these boundaries relate to lines in the data set? How do the streaks in the reconstruction relate to the data set?

Region of Interest (ROI) Data

In region of interest tomography, one chooses a subset of the object, called a *region of interest (ROI)*, to reconstruct. ROI data consist of all lines that meet this region, and the ROI problem is to reconstruct the structure of the ROI from these data. Such data are also called interior data (and the interior problem). ROI CT is important in the CT of small parts of objects, so-called micro-CT [18, p. 460]. Other algorithms in ROI CT include [104] (when one knows the value of a function in part of the interior) and [105] (when the density is piecewise constant in the ROI), as well as algorithms including [52]. A singular value decomposition was developed for this problem in [63].

ROI CT is useful for medical CT and industrial nondestructive evaluation in which one is interested only in a small region of interest in an object, not the entire object. An advantage for medical applications is that ROI data gives less radiation than with complete data.

Lambda tomography [17], [18] is one important algorithm for ROI CT which will be described in section "Filtered Backprojection (FBP) for the X-Ray Transform," and the ROI reconstruction presented here uses this algorithm. The data are severely limited – they include only lines near the disk and the ROI transform is not injective (see Theorem 6), so why do the reconstructions look so good?

Limited Angle Region of Interest Tomography

In this modality data are given over lines in a limited angular range and that are restricted to pass through a given ROI. It comes up in single axis tilt electron microscopy (ET) (see Öktem's chapter in this book [71]). However, in general, ET is better understood as a three-dimensional problem, and this will be presented in the next paragraph.

Electron Microscope Tomography (ET) Over Arbitrary Curves

Now a full three-dimensional problem, electron microscope tomography (ET), is considered. The notation is as in Öktem's chapter in this book [71], which has detailed information about the physics, biology, model, and mathematics of ET. A reconstruction of a simple 3D phantom – the union of the following disks: with center $(0, 0, 0)$, radius $1/2$; center $(0, 0, 1)$, radius $1/2$; center $(1, -1, 1)$, radius $1/4$;

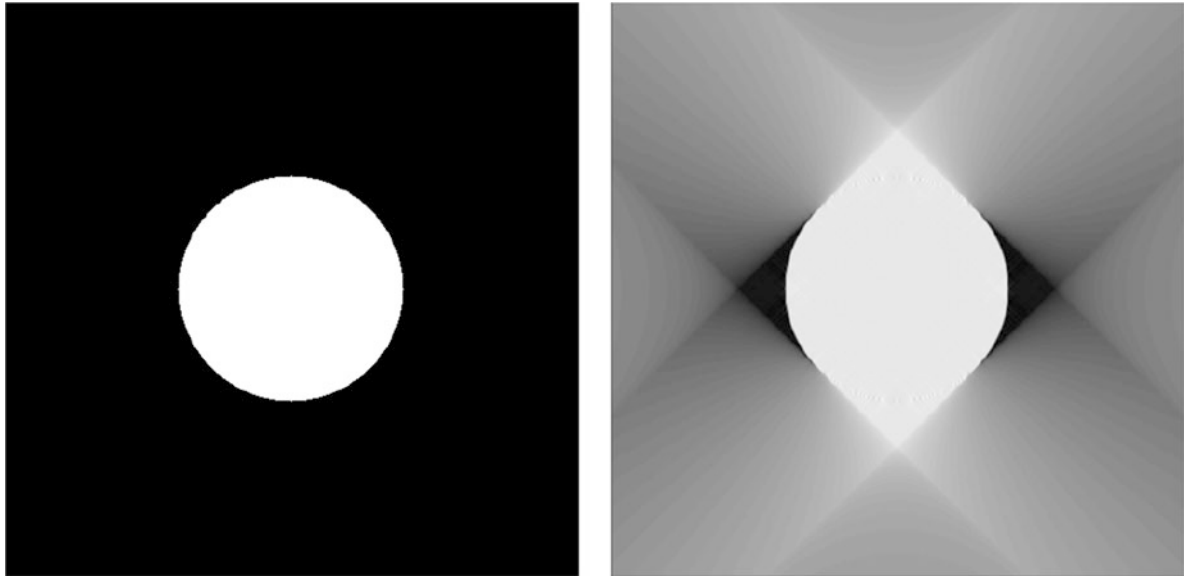


Fig. 3 Limited angle reconstruction of a disk. Original image (*left*) and reconstruction (*right*) from a truncated filtered backprojection (FBP) algorithm (26) using data in the angular range, $\varphi \in [-\pi/4, \pi/4]$. Note the streak artifacts and the missing boundaries in the limited angle reconstructions [26, ©IOP Publishing. Reproduced by permission of IOP Publishing. All rights reserved]

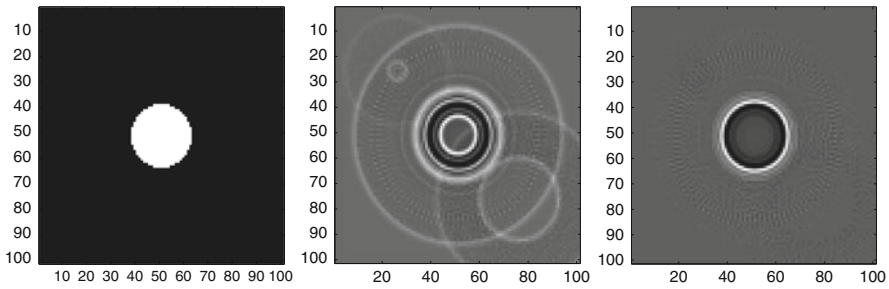


Fig. 4 Reconstruction from conical tilt data. Cross section with the $x - y$ plane of the phantom described in this section (*left*), $\mathcal{L}_\Delta$ reconstruction (*center*, see Eq. (28)) and $\mathcal{L}_S$ reconstruction (*right*, see Eq. (29)). The center of the cross section is the origin, and the range in x and y is from -2 to 2. This research was done with Sohhyun Chung and Tania Bakhos while they were Tufts undergraduates [82, ©Scuola Normale Superiore, Pisa. Reproduced by permission with Scuola Normale Superiore, Pisa, all rights reserved]

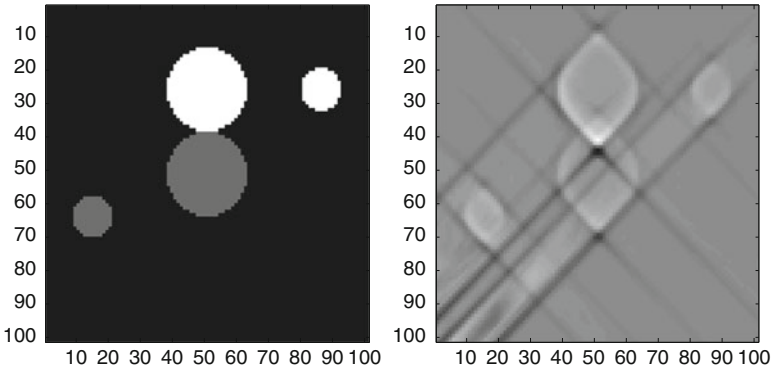


Fig. 5 Reconstruction from conical tilt data. Cross section of phantom in the plane $x = -y$ (*left*) and $\mathcal{L}_\Delta$ reconstruction in that plane (*right*). The $x - y$ plane cuts the picture in half with a *horizontal line* [82, ©Scuola Normale Superiore, Pisa. Reproduced by permission with Scuola Normale Superiore, Pisa, all rights reserved]

and center $(-1, 1, -1/2)$, radius $1/4$ – is analyzed. The disks above the $x - y$ plane have density two and the others have density one.

Conical tilt ET data is described in section "Algorithms in Conical Tilt ET" and Öktem's chapter in this book [71]. In this example, line integrals are given over all lines in space with angle $\alpha = \pi/4$ with the $z-$axis. Reconstructions are given from two algorithms that are described in section "Electron Microscope Tomography (ET) Over Arbitrary Curves." The operators are $\mathcal{L}_\Delta$ (given in Eq. (28)) and $\mathcal{L}_S$ (given in Eq. (29)).

Artifacts are added in the $\mathcal{L}_\Delta$ reconstruction in Fig. 4 and in Fig. 5 which shows the plane containing the centers of the disks and the $z-$axis (axis of rotation of the scanner). These figures are remarkable because the $\mathcal{L}_\Delta$ reconstruction has so

many added artifacts compared to the $\mathcal{L}_S$ reconstruction, although these operators are not very different (see section "Algorithms in Conical Tilt ET"). Why are the reconstructions so different?

Reconstructions of real specimens from single axis tilt data show some of the same strengths and limitations (see, e.g., [80, 83] and Öktem's chapter in this book [71]). However, the added artifacts have different properties, and since the data are so noisy, other factors affect reconstructions.

Synthetic-Aperture Radar Imaging

In synthetic-aperture radar (SAR) imaging, a region of interest on the surface of the earth is illuminated by electromagnetic waves from an airborne platform such as a plane or satellite. For more detailed information on SAR imaging, including several open problems in SAR imaging, the reader is referred to [7, 8] and to the chapter in this handbook by Cheney and Borden [9]. The backscattered waves are picked up at a receiver or receivers, and the goal is to reconstruct an image of the region based on such measurements. In monostatic SAR, the transmitter and receiver are located on the same platform. In bistatic SAR, the transmitter and receiver are on independently moving trajectories.

While monostatic SAR imaging is the one that is widely used, bistatic SAR imaging offers several advantages. The receivers in comparison to transmitters are not active sources of electromagnetic radiation and hence are more difficult to detect if flown in an unsafe environment. Since the transmitter and receiver are at different points in space, bistatic SAR systems are more resistant to electronic countermeasures such as target shaping to reduce scattering in the direction of incident waves [48].

The reconstruction of the image based on the measurement of the backscattered waves is in general a hard problem. However, ignoring contributions of multiply backscattered waves linearizes the relation between the image to be recovered and the backscattered waves and is easier to analyze. Due to this reason, a linearizing approximation called the Born approximation that ignores contribution from multiply scattered waves is widely used in SAR image reconstruction.

The Linearized Model in SAR Imaging

Let $\gamma_T(s)$ and $\gamma_R(s)$ for $s \in (s_0, s_1)$ be the trajectories of the transmitter and receiver, respectively. The propagation of electromagnetic waves can be described by the scalar wave equation

$$\left(\Delta - \frac{1}{c^2}\partial_t^2\right) E(x,t) = -P(t)\delta(x - \gamma_T(s)), \tag{5}$$

where c is the speed of electromagnetic waves in the medium, $E(x,t)$ is each component of the electric field, and $P(t)$ is the transmit waveform sent to the transmitter antenna. The wave speed c is spatially varying due to inhomogeneities

present in the medium, and one can assume that it is a perturbation of the constant background speed of propagation c_0 of the form

$$\frac{1}{c^2(x)} = \frac{1}{c_0^2} + \tilde{V}(x).$$

One assumes that $\tilde{V}(x)$ only varies over a two-dimensional surface: the surface of the earth. Therefore, $\tilde{V}$ can be represented as a function of the form

$$\tilde{V}(x) = V(x)\delta_0(x_3)$$

where it will be assumed that the earth's surface is represented by the $x = (x_1, x_2)$ plane. The background Green's function g is the solution of the following equation:

$$\left(\Delta - \frac{1}{c_0^2}\partial_t^2\right) g(x,t) = -\delta_0(x)\delta_0(t).$$

This is given by

$$g(x,t) = \frac{\delta(t - \|x\|/c_0)}{4\pi \|x\|}. \tag{6}$$

Now the incident field E^{in} due to the source $s(x,t) = P(t)\delta(x - \gamma_T(s))$ is

$$\begin{aligned} E^{\text{in}}(x,t) &= \int g(x-y, t-\tau)s(y,\tau)\mathrm{d}y\mathrm{d}\tau \\ &= \frac{P(t - \|x - \gamma_T(s)\|/c_0)}{4\pi \|x - \gamma_T(s)\|}. \end{aligned}$$

Let E denote the total field of the medium, $E = E^{\text{in}} + E^{\text{sc}}$, where E^{sc} is the scattered field. This can be written using the Lippmann-Schwinger equation:

$$E^{\text{sc}}(z,t) = \int g(z-x, t-\tau)\partial_t^2 E(x,\tau)V(x)\mathrm{d}x\mathrm{d}\tau. \tag{7}$$

This equation is linearized by replacing the total field E on the right-hand side of the above equation by E^{in}. This is known as the Born approximation. The linearized scattered wave field $E^{\text{sc}}_{\text{lin}}(\gamma_R(s), t)$ at the receiver location $\gamma_R(s)$ is then

$$E^{\text{sc}}_{\text{lin}}(\gamma_R(s), t) = \int g(x - \gamma_R(s), t-\tau)\partial_t^2 E^{\text{in}}(x,\tau)V(x)\mathrm{d}x\mathrm{d}\tau.$$

Substituting the expression for E^{in} into this equation and integrating, one obtains the following expression for the linearized scattered wave field:

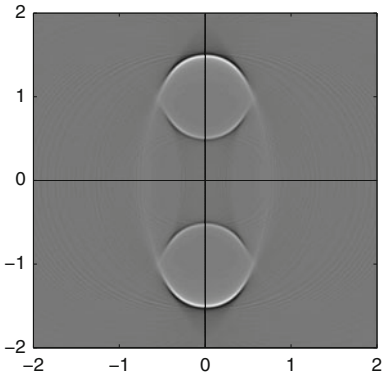


Fig. 6 Reconstruction of a disk centered on the positive y-axis from integrals over ellipses (with constant distance between the foci) centered on the x-axis and with foci in $[-3,3]$. Notice that some boundaries of the disk are missing, and there is a copy of the disk below the axis. This was originally from the Tufts University Senior Honors thesis of Howard Levinson and published in [55, Reproduced with kind permission from Springer Science+Business Media: © Springer Verlag]

$$E^{\mathrm{sc}}_{\mathrm{lin}}(\gamma_R(s),t) = \int e^{-\mathrm{i}\omega(t-\frac{1}{c_0}R(s,x))} A(s,x,\omega)V(x)\mathrm{d}x\mathrm{d}\omega, \tag{8}$$

where

$$R(s,x) = \|\gamma_T(s) - x\| + \|x - \gamma_R(s)\|$$

and

$$A(s,x,\omega) = \omega^2 p(\omega)((4\pi)^2 \|\gamma_T(s) - x\| \, \|\gamma_R(s) - x\|)^{-1}, \tag{9}$$

where p is the Fourier transform of P. The function A includes terms that take into account the transmitted waveform and geometric spreading factors. The inverse of the norms appears in A due to the background Green's function, (6).

The reconstruction in Fig. 6 of a disk centered on the positive y-axis from integrals of it over ellipses with foci moving along the x-axis offset by a constant distance (which is simplified model of (8)) highlights some of the features in SAR image reconstruction. Some part of the boundary is not stably reconstructed, and an artifact of the true image appears as a reflection about the x-axis along with streak artifacts. Looking at the reconstructed image, one sees that at least visually, the created artifact is as strong as the true image. Microlocal analysis of the operators appearing in SAR imaging will make precise and justify all these observations. This will be addressed in section "SAR Imaging."

General Observations

In each reconstruction in this section, some object boundaries are visible and others are not. In fact, if one looks more carefully at the reconstructions, one can notice that in each case, the only feature boundaries that are clearly defined are those tangent to lines in the data set for the problem. Example 1 illustrates this in a naive way: one sees singularities in the Radon data exactly when the lines of integration are tangent to the boundary of the object. The goal of this chapter is to make the idea mathematically rigorous.

The conical tilt ET reconstructions in section "Electron Microscope Tomography (ET) Over Arbitrary Curves" have artifacts if one uses a certain algorithm but apparently not when one uses another similar one. The reconstruction related to radar in Fig. 6 has an artifact that is a reflected image of the disk.

In Sect. 4, deep mathematical ideas from microlocal analysis will be introduced to classify singularities and what certain operators do to them. In Sect. 5 these microlocal ideas will be used to explain the visible and invisible singularities for limited data X-ray CT as well as the added singularities in ET and radar.

3 Properties of Tomographic Transforms

In this section, after introducing some functional analysis, basic properties of transforms in X-ray tomography and electron microscope tomography are presented. The microlocal properties of radar will be given later in section "SAR Imaging."

Function Spaces

The open disk in $\mathbb{R}^2$ centered at the origin and of radius $r > 0$ will be denoted $D(r)$.

The set $C^\infty(\mathbb{R}^n)$ consists of all smooth functions on $\mathbb{R}^n$, that is, functions that are continuous along with their derivatives of all orders, and $\mathcal{D}(\mathbb{R}^n)$ is the set of smooth functions of compact support. Its dual space – the set of all continuous linear functionals on $\mathcal{D}(\mathbb{R}^n)$ (given the weak-* topology) – is denoted $\mathcal{D}'(\mathbb{R}^n)$ and is called the *set of distributions*. If u is a locally integrable function, then u is a distribution with the standard definition

$$\langle u, f \rangle = u(f) = \int_{\mathbb{R}^n} u(x) f(x) \, \mathrm{d}x$$

for $f \in \mathcal{D}(\mathbb{R}^n)$ since $u(x) f(x)$ is an integrable function of compact support.

If Ω is an open set in $\mathbb{R}^n$, then $\mathcal{D}(\Omega)$ is the set of smooth functions compactly supported in Ω. Its dual space with the weak-* topology is denoted $\mathcal{D}'(\Omega)$.

The *Schwartz space* of rapidly decreasing functions is the set $\mathcal{S}(\mathbb{R}^n)$ of all smooth functions that decrease (along with all their derivatives) faster than any power of $1/\|x\|$ at infinity. Its dual space, $\mathcal{S}'(\mathbb{R}^n)$, is the set of all continuous linear

functionals on $\mathcal{S}(\mathbb{R}^n)$ with the weak-* topology (convergence is pointwise: $u_k \to u$ in $\mathcal{S}'(\mathbb{R}^n)$ if, for each $f \in \mathcal{S}(\mathbb{R}^n)$, $u_k(f) \to u(f)$). They are called *tempered distributions*. Any function that is measurable and bounded above by some power of $(1 + \|x\|)$ is in $\mathcal{S}'(\mathbb{R}^n)$ since its product with any Schwartz function is integrable.

A distribution u is supported in the closed set K if for all functions $f \in \mathcal{D}(\mathbb{R}^n)$ with support disjoint from K, $u(f) = 0$. The support of u, supp(u), is the smallest closed set in which u is supported.

Example 2. The *Dirac delta function at zero* is an important distribution that is not a function. It is defined $\langle \delta_0, f \rangle = \delta_0(f) = f(0)$. Note that if f is supported away from the origin, then $\delta_0(f) = 0$ since $f(0) = 0$. Therefore, the Dirac delta function has support $\{0\}$.

Let $\mathcal{E}'(\mathbb{R}^n)$ be the set of distributions that have compact support in $\mathbb{R}^n$. If Ω is an open set in $\mathbb{R}^n$, then $\mathcal{E}'(\Omega)$ is the set of distributions with compact support contained in Ω. For example, on the real line, $\delta \in \mathcal{E}'((-1, 1))$.

If $f \in L^1(\mathbb{R}^n)$, then its Fourier transform and inverse are

$$\begin{aligned} \mathcal{F}f(\xi) = \hat{f}(\xi) &= \frac{1}{(2\pi)^{n/2}} \int_{x \in \mathbb{R}^n} e^{-\mathrm{i}x \cdot \xi} f(x)\, \mathrm{d}x \\ \mathcal{F}^{-1} f(x) = \check{f}(x) &= \frac{1}{(2\pi)^{n/2}} \int_{\xi \in \mathbb{R}^n} e^{\mathrm{i}x \cdot \xi} f(\xi)\, \mathrm{d}\xi\,. \end{aligned} \tag{10}$$

The Fourier transform is linear and continuous from $L^1(\mathbb{R}^n)$ to the space of continuous functions that converge to zero at ∞. Furthermore, $\mathcal{F}$ is an isomorphism on $L^2(\mathbb{R}^n)$ and an isomorphism on $\mathcal{S}(\mathbb{R}^n)$ and, therefore, on $\mathcal{S}'(\mathbb{R}^n)$. More information about these topics can be found in [86], for example.

Basic Properties of the Radon Line Transform

In this section fundamental properties of the Radon line transform, $\mathcal{R}_L$, are derived, see [66]. This will provide a connection between the transforms and microlocal analysis in Sect. 4.

Theorem 1 (General Projection Slice Theorem). *Let $f \in L^1(\mathbb{R}^2)$. Now let $h \in L^\infty(\mathbb{R})$ and $\omega \in S^1$. Then,*

$$\int_{x \in \mathbb{R}^2} f(x) h(x \cdot \omega)\, \mathrm{d}x = \int_{p=-\infty}^{\infty} \mathcal{R}_L f(\omega, p) h(p)\, \mathrm{d}p. \tag{11}$$

Proof. Let $\omega \in S^1$. First, note that the function $x \mapsto f(x)h(x \cdot \omega)$ is in $L^1(\mathbb{R}^2)$ since h is bounded and measurable. For the same reason, the function

$$(p,t) \mapsto f(p\omega + t\omega^{\perp})h(p)$$

is in $L^1(\mathbb{R}^2)$. Therefore,

$$\int_{x\in\mathbb{R}^2} f(x)h(x\cdot\omega)\,\mathrm{d}x = \int_{p=-\infty}^{\infty}\int_{t=-\infty}^{\infty} f(p\omega + t\omega^{\perp})h\,(p)\;\mathrm{d}t\,\mathrm{d}p \tag{12}$$

$$= \int_{p=-\infty}^{\infty} \mathcal{R}_L f(\omega, p)h(p)\,\mathrm{d}p \tag{13}$$

where (12) holds by rotation invariance of the Lebesgue integral and then Fubini's theorem and since $p = \omega\cdot(p\omega + t\omega^{\perp})$. The equality (13) holds by the definition of $\mathcal{R}_L$. □

Let $\mathcal{S}(S^1\times\mathbb{R})$ be the set of smooth functions on $S^1\times\mathbb{R}$ that decrease (along with all their derivatives) faster than any power of $1/|p|$ at infinity uniformly in ω, and let $\mathcal{S}'(S^1\times\mathbb{R})$ be its dual. The *partial Fourier transform* is defined for $g\in L^1(S^1\times\mathbb{R})$ as

$$\mathcal{F}_p g(\omega,\tau) = \frac{1}{\sqrt{2\pi}}\int_{p\in\mathbb{R}} e^{-\mathrm{i}p\tau} g(\omega,\tau)\,\mathrm{d}\tau\,. \tag{14}$$

Because the Fourier transform is an isomorphism on $\mathcal{S}(\mathbb{R})$, this transform and its inverse are defined and continuous on $\mathcal{S}'(S^1\times\mathbb{R})$.

The *Fourier Slice Theorem* is an important corollary of Theorem 1.

Theorem 2 (Fourier Slice Theorem). *Let $f\in L^1(\mathbb{R}^2)$. Then for $(\omega,\tau)\in S^1\times\mathbb{R}$,*

$$\mathcal{F}f(\tau\omega) = \frac{1}{\sqrt{2\pi}}\mathcal{F}_p\mathcal{R}f(\omega,\tau)\,.$$

To prove this theorem, one applies the General Projection Slice Theorem 1 to the function $h(p) = e^{-\mathrm{i}p\tau}$.

The Fourier Slice Theorem provides a proof that $\mathcal{R}_L$ is invertible on domain $L^1(\mathbb{R}^2)$ since $\mathcal{F}_p$ is invertible on domain $L^1(S^1\times\mathbb{R})$. Zalcman constructed a nonzero function that is integrable on every line in the plane and whose line transform is identically zero [107]. Of course, his function is not in $L^1(\mathbb{R}^2)$.

This theorem also provides a proof of invertibility for the limited angle problem.

Theorem 3 (Limited Angle Theorem). *Let $f\in\mathcal{E}'(\mathbb{R}^2)$ and let $\Phi\in(0,\pi/2)$. If $\mathcal{R}_L f(\omega(\varphi),p) = 0$ for $\varphi\in(-\Phi,\Phi)$ and all p, then $f = 0$.*

However, there are nonzero functions $f\in\mathcal{S}(\mathbb{R}^2)$ with $\mathcal{R}_L f(\omega(\varphi),p) = 0$ for $\varphi\in(-\Phi,\Phi)$ and all p.

Proof. Let $f \in \mathcal{E}'(\mathbb{R}^2)$ and assume $\mathcal{R}_L f(\omega(\varphi), p) = 0$ for $\varphi \in (-\Phi, \Phi)$ and all p. By the Fourier Slice Theorem, which is true for $\mathcal{E}'(\mathbb{R}^2)$ [45],

$$\mathcal{F}f(\tau\omega(\varphi)) = \frac{1}{\sqrt{2\pi}}\mathcal{F}_p\mathcal{R}_L f(\omega(\varphi), \tau) = 0 \quad \text{for } \varphi \in (-\Phi, \Phi), \tau \in \mathbb{R} \tag{15}$$

and this expression is zero because $\mathcal{R}_L f(\omega(\varphi), \tau) = 0$ for such (φ, τ). This shows that $\mathcal{F}f$ is zero on the open cone

$$V = \{\tau\omega(\varphi) : \tau \neq 0, \varphi \in (-\Phi, \Phi)\}.$$

Since f has compact support, $\mathcal{F}f$ is real analytic, and so $\mathcal{F}f$ must be zero everywhere since it is zero on the open set V. This shows $f = 0$.

To prove the second part of the theorem, let $\tilde{f}$ be any nonzero Schwartz function supported in the cone V and let $f = \mathcal{F}^{-1}\left(\tilde{f}\right)$. Since $\tilde{f}$ is nonzero and in $\mathcal{S}(\mathbb{R}^2)$, so is f. Using (15) but starting with $\mathcal{F}f = 0$ in V, one sees that $\mathcal{R}_L f$ is zero in the limited angular range. □

Another application of these theorems is the classical Range Theorem for this transform.

Theorem 4 (Range Theorem [28,41]). *Let $g \in \mathcal{S}(S^1 \times \mathbb{R})$. Then g is in the range of $\mathcal{R}_L$ on domain $\mathcal{S}(\mathbb{R}^2)$ if and only if*

1. *$g(\omega, p) = g(-\omega, -p)$ for all $(\omega, p) \in S^1 \times \mathbb{R}$*
2. *For each $m \in \{0, 1, 2, \ldots\}$, $\int_{p\in\mathbb{R}} g(\omega, p)p^m \, \mathrm{d}p$ is a polynomial in $\omega \in S^1$ that is homogeneous of degree m.*

Proof (Sketch). The necessary part of the theorem follows by applying the General Projection Slice Theorem to $h(p) = p^m$ for m a nonnegative integer:

$$\int_{p\in\mathbb{R}} \mathcal{R}_L f(\omega, p)p^m \, \mathrm{d}p = \int_{x\in\mathbb{R}^2} f(x)(x \cdot \omega)^m \, \mathrm{d}x$$

and after multiplying out $(x \cdot \omega)^m$ in the coordinates of ω, one sees that the right-hand integral is a polynomial in these coordinates homogeneous of degree m. The sufficiency part is much more difficult to prove. One uses the Fourier Slice Theorem to construct a function f satisfying $\mathcal{F}f(\tau\omega) = \frac{1}{\sqrt{2\pi}}\mathcal{F}_p g(\omega, \tau)$. Since $\mathcal{F}_p g$ is smooth and rapidly decreasing in p, $\mathcal{F}f$ is smooth away from the origin and rapidly decreasing in x. The subtle part of the proof in [41] is to show $\mathcal{F}f$ is smooth at the origin, and this is done using careful estimates on derivatives using the moment conditions, (2) of Theorem 4. Once that is known, one can conclude $\mathcal{F}f \in \mathcal{S}(\mathbb{R}^2)$ and so $f \in \mathcal{S}(\mathbb{R}^2)$. □

The support theorem for $\mathcal{R}_L$ is elegant and has motivated a large range of generalizations such as [5, 6, 42, 53, 57, 78].

Theorem 5 (Support Theorem [10, 28, 41]). *Let f be a distribution of compact support (or a function in $\mathcal{S}(\mathbb{R}^2)$) and let $r > 0$. Assume $\mathcal{R}_L f$ is zero for all lines that are disjoint from the disk $D(r)$. Then* $\operatorname{supp}(f) \subset D(r)$.

This theorem implies that the exterior problem has a unique solution; in this case, $D(r)$ is the excluded region. The proof is tangential to the main topics of this chapter, and it can be found in [10, 28, 41, 43, 92].

Counterexamples to the support theorem exist for functions that do not decrease rapidly at ∞ (e.g., [43, 106] or the singular value decompositions in [73, 76]).

A corollary of these theorems shows that exact reconstruction is impossible from ROI data where $D(r)$ is the disk centered at the origin in $\mathbb{R}^2$ and of radius $r > 0$.

Theorem 6. *Consider the ROI problem with region of interest the unit disk $D(1)$. Let $r \in (1, \infty)$. Then there is a function $f \in \mathcal{D}(D(r))$ that is not identically zero in $D(1)$ but for which $\mathcal{R}_L f$ is zero for all lines that intersect $D(1)$.*

Proof (Sketch). Let $h(p)$ be a smooth nonzero nonnegative function supported in $(1, r)$ and let $g(\omega, p) = h(|p|)$. Since g is independent of ω, the moment conditions from the Range Theorem (Theorem 4), are trivially satisfied, so that theorem shows that there is a function $f \in \mathcal{S}(\mathbb{R}^2)$ with $\mathcal{R}_L f = g$. By the support theorem, f is supported in the disk $D(r)$. To show f is nonzero in the ROI, $D(1)$, one uses [10, p. 2725, equation (18)]. This is also proven in [66, p. 169, VI.4], and Natterer shows that such null functions do not oscillate much in the ROI. It will be shown in Sect. 5 that null functions are smooth in the ROI, too. □

Continuity Results for the X-Ray Transform

In this section some basic continuity theorems for $\mathcal{R}_L$ are presented.

A simple proof shows that $\mathcal{R}_L$ is continuous from $C_c(D(M))$ to $C_c(S_M)$ where we define $S_M = S^1 \times [-M, M]$. First, one uses uniform continuity of f to show $\mathcal{R}_L f$ is a continuous function. Then, the proof that $\mathcal{R}_L$ is continuous is based on the estimate

$$|\mathcal{R}_L f(\omega, p)| \leq \pi M^2 \|f\|_\infty$$

where $\|f\|_\infty$ is the (essential) supremum norm of f. A stronger theorem has been proven by Helgason.

Theorem 7 ([41]). $\mathcal{R}_L : \mathcal{S}(\mathbb{R}^2) \to \mathcal{S}(S^1 \times \mathbb{R})$ *is continuous.*

The proof of the next theorem follows from the calculations in the proof of the General Projection Slice Theorem.

Theorem 8. $\mathcal{R}_L : L^1(\mathbb{R}^2) \to L^1(S^1 \times \mathbb{R})$ *is continuous.*

Proof. By taking absolute values in (11) with $h = 1$ and then integrating with respect to ω, one sees that $\|f\|_{L^1(\mathbb{R}^2)} \geq (2\pi)\,\|\mathcal{R}_L f\|_{L^1(S^1\times\mathbb{R})}$ and so $\mathcal{R}_L$ is continuous on L^1. □

Continuity results for $\mathcal{R}_L$ in Sobolev spaces were given in [40, 45, 59] for functions of fixed compact support.

Filtered Backprojection (FBP) for the X-Ray Transform

To state the most commonly used inversion formula, filtered backprojection, one first defines the *dual line transform*. For $g \in L^1(S^1 \times \mathbb{R})$ and $x \in \mathbb{R}^2$,

$$\mathcal{R}_L^* g(x) = \int_{\omega\in S^1} g(\omega, x\cdot\omega)\,\mathrm{d}\omega. \tag{16}$$

For each $\omega \in S^1$, $x \in \ell(\omega, x\cdot\omega)$, so $\mathcal{R}_L^* g(x)$ is the integral of g over all lines through x. The transform $\mathcal{R}_L^*$ is the formal dual to $\mathcal{R}_L$ in the sense that for $f \in \mathcal{S}(\mathbb{R}^2)$ and $g \in \mathcal{S}(S^1 \times \mathbb{R})$,

$$\langle \mathcal{R}_L f, g\rangle_{L^2(S^1\times\mathbb{R})} = \langle f, \mathcal{R}_L^* g\rangle_{L^2(\mathbb{R}^2)}\ .$$

Because $\mathcal{R}_L : \mathcal{S}(\mathbb{R}^2) \to \mathcal{S}(S^1 \times \mathbb{R})$ is continuous, $\mathcal{R}_L^* : \mathcal{S}'(S^1 \times \mathbb{R}) \to \mathcal{S}'(\mathbb{R}^2)$ is weakly continuous.

The Lambda operator is defined on functions $g \in \mathcal{S}(S^1 \times \mathbb{R})$ by

$$\Lambda_p g(\omega, p) = \mathcal{F}_p^{-1}\left(|\tau|\left(\mathcal{F}_p g(\omega, \cdot)\right)\right). \tag{17}$$

Theorem 9 (Filtered Backprojection (FBP) [66, 85, 87]). *Let $f \in \mathcal{S}(\mathbb{R}^2)$. Then,*

$$f = \frac{1}{4\pi}\mathcal{R}_L^* \Lambda_p \mathcal{R}_L f. \tag{18}$$

This formula is valid for $f \in \mathcal{E}'(\mathbb{R}^2)$.

Filtered backprojection is an efficient, fast reconstruction method that is easily implemented [67] by using an approximation to the operator Λ_p that is convolution with a function (see, e.g., [66] or [87]). Note that FBP requires data over all lines

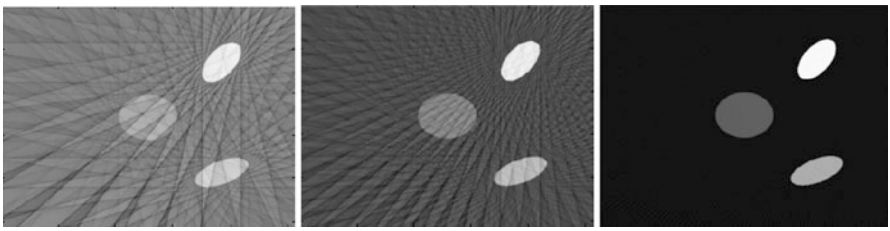


Fig. 7 FBP reconstructions of phantom consisting of three ellipses. The *left* reconstruction uses 18 angles, the *middle* 36 angles, and the *right* one 180 angles

through the object – it is not local: in order to find $f(x)$, one needs data $\mathcal{R}_L f$ over all lines in order to evaluate $\Lambda_p \mathcal{R}_L f$ (which involves a Fourier transform).

To see the how sensitive FBP is to the number of the angles used, reconstructions are provided in Fig. 7 using 18, 36, and 180 angles. One can see that using too few angles creates artifacts. An optimal choice of angles and values of p can be determined using sampling theory [15, 16, 66].

Proof (Proof of Theorem 9). Let $f \in \mathcal{S}(\mathbb{R}^2)$. First, one writes the two-dimensional Fourier inversion formula in polar coordinates:

$$f(x) = \frac{1}{2(2\pi)} \int_{\omega \in S^1} \int_{\tau \in \mathbb{R}} e^{\mathrm{i}x \cdot (\tau\omega)} \hat{f}(\tau\omega) |\tau| \, \mathrm{d}\tau \, \mathrm{d}\omega \tag{19}$$

$$= \frac{1}{4\pi} \int_{\omega \in S^1} \int_{\tau \in \mathbb{R}} \frac{e^{\mathrm{i}\tau(\omega \cdot x)}}{\sqrt{2\pi}} |\tau| \left(\mathcal{F}_p \mathcal{R}_L f\right)(\omega, \tau) \, \mathrm{d}\tau \, \mathrm{d}\omega \tag{20}$$

$$= \frac{1}{4\pi} \int_{\omega \in S^1} \left(\Lambda_p \mathcal{R}_L f\right)(\omega, \omega \cdot x) \, \mathrm{d}\omega = \frac{1}{4\pi} \mathcal{R}_L^* \Lambda_p \mathcal{R}_L f(x) \,. \tag{21}$$

The factor of $1/2$ in front of the integral in (19) occurs because the integral is over $\tau \in \mathbb{R}$ rather than $\tau \in [0, \infty)$. The Fourier Slice Theorem (Theorem 2) is used in (20), and the definitions of Λ_p and of $\mathcal{R}_L^*$ are used in (21). The integrals above exist because f, $\mathcal{F} f$, and $\mathcal{R}_L f$ are all rapidly decreasing at infinity.

The proof that the FBP formula is valid for $f \in \mathcal{E}'(\mathbb{R}^2)$ will now be given. Since $f \in \mathcal{E}'(\mathbb{R}^2)$, the Fourier Slice Theorem holds for f [45], and

$$\mathcal{F}_p \mathcal{R}_L f(\omega, \tau) = \sqrt{2\pi} \mathcal{F} f(\tau\omega)$$

is a smooth function that is polynomially increasing [86]. So, $|\tau| \mathcal{F}_p \mathcal{R}_L(\omega, \tau)$ is a polynomially increasing continuous function and therefore in $\mathcal{S}'(S^1 \times \mathbb{R})$. Since the inverse Fourier transform, $\mathcal{F}_p^{-1}$, maps $\mathcal{S}'$ to $\mathcal{S}'$, $\Lambda_p \mathcal{R}_L f$ is a distribution in $\mathcal{S}'(S^1 \times \mathbb{R})$. By duality with $\mathcal{S}$, $\mathcal{R}_L^* : \mathcal{S}'(S^1 \times \mathbb{R}) \to \mathcal{S}'(\mathbb{R}^2)$, so $\mathcal{R}_L^* \Lambda_p \mathcal{R}_L f$ is defined for $f \in \mathcal{E}'(\mathbb{R}^2)$. Since the Fourier Slice Theorem holds for f [45], the FBP formula can be proved for f as is done above for $\mathcal{S}$ (see [26] more generally). □

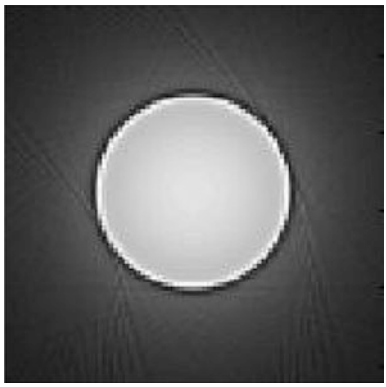


Fig. 8 ROI reconstruction by Tufts undergraduate Stephen Bidwell from simulated data for the characteristic function of a *circle* using the operator $\mathcal{L}_{x,\mu}$ given in (23) [4, ©Tufts University]

Limited Data Algorithms

In limited data problems, some data are missing, and reconstruction methods are now presented for ROI CT, limited angle CT, and limited angle ROI CT.

ROI Tomography

Lambda tomography [17,18,96] is an effective and easy-to-implement algorithm for ROI CT. The fundamental idea is to replace Λ_p by $-d^2/dp^2$ in the FBP formula. The relation between these two operators is that $\Lambda_p^2 = -d^2/dp^2$, which will be justified in Example 9. This motivates the definition

$$\mathcal{L}_x f := \frac{1}{4\pi}\mathcal{R}_L^*\left(-\frac{d^2}{dp^2}\mathcal{R}_L f\right). \tag{22}$$

FBP is not local; to calculate $\Lambda_p \mathcal{R}_L f(\omega, p)$, one needs data over all lines to take the Fourier transform $\mathcal{F}_p$. Therefore, one needs data over all lines through the object to calculate f using FBP (18). This means that FBP cannot be used on ROI data.

The advantage of Lambda tomography over FBP is that $\mathcal{L}_x$ is local in the following sense. To calculate $\mathcal{L}_x f$, one needs the values of $\left(-d^2/dp^2\right)\mathcal{R}_L f$ at all lines through x (since $\mathcal{R}_L^*$ evaluated at x integrates over all such lines). Furthermore, $\left(-d^2/dp^2\right)$ is a local operator, and to calculate $\left(-d^2/dp^2\right)\mathcal{R}_L f$ at a line through x, one needs only data $\mathcal{R}_L f$ over lines close to x. Therefore, one needs only data over all lines near x to calculate $\mathcal{L}_x f$. Therefore, $\mathcal{L}_x$ can be used on ROI data. Although Lambda CT reconstructs $\mathcal{L}_x f$, not f itself, it shows boundaries very clearly (see Fig. 8 and [17]).

Kennan Smith developed an improved local operator that shows contours of objects, not just boundaries. His idea was to add a positive multiple of $\mathcal{R}_L^*\mathcal{R}_L f$ to the reconstruction to get

$$\mathcal{L}_{x,\mu} f = \frac{1}{4\pi} \mathcal{R}_L^* \left(\left(-\frac{d^2}{dp^2} + \mu \right) \mathcal{R}_L f \right) \tag{23}$$

for some $\mu > 0$. Using (35), one sees that

$$\mathcal{R}_L^* \left(\mu \mathcal{R}_L f \right)(x) = \left(\frac{2\mu}{\|x\|} * f \right)(x), \tag{24}$$

so this factor adds contour to the reconstruction since the convolution with $2\mu/\|x\|$ emphasizes the values of f near x. Lambda reconstructions look much like FBP reconstructions even though they are local. A discussion of how to choose μ to counteract a natural cupping effect of $\mathcal{L}_x$ is given in [17]. The operator $\mathcal{L}_{x,\mu}$ is local for the same reasons as $\mathcal{L}_x$ is, and it was used in the ROI reconstruction in Fig. 8.

The ideas behind Lambda CT can be adapted to a range of limited data problems including limited angle tomography(e.g., [49, 56]), exterior tomography [79], and three-dimensional problems such as cone beam CT [2, 25, 51, 103] and conical tilt electron microscopy [23]. Here is one such application.

Limited Angle CT

There are several algorithms for limited angle tomography (e.g., [3,12,49,61,104]), and two will be presented that are generalizations of FBP and Lambda CT. The key to each is to use the limited angle backprojection operator that uses angles in an interval $(-\Phi, \Phi)$ with $\Phi \in (0, \pi/2)$:

$$\mathcal{R}_{L,\lim}^* g(x) = \int_{\varphi=-\Phi}^{\Phi} g(\omega(\varphi), x \cdot \omega(\varphi)) \mathrm{d}\varphi \,. \tag{25}$$

The limited angle FBP and limited angle Lambda algorithms are

$$\mathcal{R}_{L,\lim}^* \Lambda_p \mathcal{R}_L f \quad \text{and} \quad \mathcal{R}_{L,\lim}^* \left(-\frac{d^2}{dp^2} \right) \mathcal{R}_L f \tag{26}$$

respectively. The objects in Fig. 3 are reconstructed using this limited angle FBP algorithm. Limited angle Lambda CT is local, so it can be used for the limited angle ROI data in electron microscope tomography [80, 83].

Fan Beam and Cone Beam CT

The parallel beam parameterization of lines in the plane used above is more convenient mathematically, but modern CT scanners use a single X-ray source that emits X-rays in a fan or cone beam. The source and detectors (on the other side of

the body) move around the body and quickly acquire data. This requires different parameterizations of lines.

The *fan beam parameterization* is used if the X-rays are collimated to one reconstruction plane. Let C be the curve of sources in the plane, typically a circle surrounding the object, and let $(\omega, \theta) \in C \times S^1$. Then,

$$L(\omega, \theta) = \{\omega + t\theta \,:\, t > 0\}$$

is the ray starting at ω in direction θ, and the *fan beam line transform* is

$$\int_{t=0}^{\infty} f(\omega + t\theta)\, \mathrm{d}t.$$

In this case the analogs of the formulas proved above are a little more complicated. For example, a Lambda-type operator can be designed by taking the negative second derivative in $\theta \in S^1$. The other formulas are similar and one can find them in [66, 90].

In cone beam tomography, the source is collimated to illuminate a cone in space, and the source moves in a curve around the body along with the detectors. This scanner images a volume in the body, rather than a planar region as $\mathcal{R}_L$, and the fan beam transform does. However, the reconstruction formulas are more subtle [50], and one can understand these difficulties using microlocal analysis [25, 30]. Related issues come up in conical tilt ET as will be described in section "Microlocal Analysis of Conical Tilt Electron Microscope Tomography (ET)."

These data acquisition methods have several advantages over parallel beam data acquisition. First, the scanners are simpler to build than the original CT scanners, which took data using the parallel geometry, and so a single X-ray source (or several collimated sources) and detector(s) were translated to get data over parallel lines in one direction, and then the source and detector were rotated to get lines for other directions. Second, they can acquire data more quickly than old style parallel beam scanners since the fan beam X-ray source and detector array move in a circle around the object.

This is all discussed in Herman's chapter in this book [44].

Algorithms in Conical Tilt ET

Conical tilt ET [108] is a new data acquisition geometry in ET that has the potential to provide faster data acquisition as well as clearer reconstructions. The algorithms used for the conical tilt ET reconstructions will be given in section "Electron Microscope Tomography (ET) Over Arbitrary Curves." This will lay the groundwork to understand why the reconstructions in that section from two very similar algorithms are so dramatically different. The model and mathematics are fully discussed in Öktem's chapter in this book [71].

First the notation is established. For $\omega \in S^2$, denote the plane through the origin perpendicular to ω by

$$\omega^P = \{x \in \mathbb{R}^3 \,:\, x \cdot \omega = 0\}. \tag{27}$$

The tangent space to the sphere S^2 is

$$T(S^2) = \{(\omega, x) \,:\, \omega \in S^2, x \in \omega^P\}$$

since the plane ω^P is the tangent plane to S^2 at ω. This gives the *parallel beam* parameterization of lines in space: for $(\omega, x) \in T(S^2)$, the line

$$L(\omega, x) = \{x + t\omega \,:\, t \in \mathbb{R}\}.$$

If ω is fixed, then the lines $L(\omega, x)$ for $x \in \omega^P$ are all parallel. As noted in Öktem's chapter in this book [71], ET data are typically taken on a curve $S \subset S^2$. This means the lines in the data set are parameterized by

$$\mathcal{M}_S = \{(\omega, x) \,:\, \omega \in S, x \in \omega^P\}.$$

So, for $f \in L^1(\mathbb{R}^3)$, the ET data of f for lines parallel to S can be modeled as the *parallel beam transform*

$$\mathcal{P}_S f(\omega, x) = \int_{t \in \mathbb{R}} f(x + t\omega)\, \mathrm{d}t \quad \text{for } (\omega, x) \in \mathcal{M}_S.$$

Its dual transform is defined for functions g on $\mathcal{M}_S$ as

$$\mathcal{P}_S^* g(x) = \int_{\omega \in S} g(\omega, x - (x \cdot \omega)\omega) \mathrm{d}\omega,$$

where $\mathrm{d}\omega$ is the arc length measure on S. This represents the integral of g over all lines through x.

In this section, conical tilt ET is considered; an angle $\alpha \in (0, \pi/2)$ is chosen, and data are taken for angles on the latitude circle:

$$S_\alpha = \{(\sin(\alpha)\cos(\varphi), \sin(\alpha)\sin(\varphi), \cos(\alpha))\} \,:\, \varphi \in [0, 2\pi]\}.$$

Let C_α be the cone with vertex at the origin and opening angle α from the vertical axis:

$$C_\alpha = \{t\omega \,:\, \omega \in S_\alpha\}.$$

Note that C_α is the cone generated by S_α.

Here are the two algorithms for which reconstructions were given in section "Electron Microscope Tomography (ET) Over Arbitrary Curves." The first operator is a generalization of one developed for cone beam CT by Louis and Maaß [62]:

$$\mathcal{L}_\Delta f = \mathcal{P}_S^*(-\Delta_S)\mathcal{P}_S f, \tag{28}$$

where Δ_S is the Laplacian on the detector plane, ω^P. The second operator is defined

$$\mathcal{L}_S f = \mathcal{P}_S^*(-\mathcal{D}_S)\mathcal{P}_S f, \tag{29}$$

where $\mathcal{D}_S$ is the second derivative on the detector plane ω^P in the tangent direction to the curve S at ω (see Öktem's chapter in this book [71]). The next theorem helps clarify these operators by writing them as convolution operators.

Theorem 10. *Let $\mathcal{P}_S$ be the conical tilt ET transform with angle $\alpha \in (0, \pi/2)$. Let $f \in \mathcal{E}'(\mathbb{R}^3)$. Then*

$$\mathcal{P}_S^*\mathcal{P}_S f = f * I = \int_{y \in C_\alpha} \frac{f(x+y)}{\|y\|} \mathrm{d}y \tag{30}$$

$$\mathcal{L}_\Delta f = (-\Delta)\,(f * I) \tag{31}$$

$$\mathcal{L}_S f = \left(-\Delta + \csc^2(\alpha)\frac{\partial^2}{\partial z^3}\right) f * I \tag{32}$$

where I is the distribution defined for $f \in \mathcal{D}(\mathbb{R}^2)$ by

$$I(f) = \int_{y \in C_\alpha} f(y)\frac{1}{\|y\|}\mathrm{d}y$$

and $\mathrm{d}y$ *is the surface area measure on the cone C_α.*

Equation (30) makes sense since $\mathcal{P}_S^*$ integrates $\mathcal{P}_S f$ over all lines in the data set through x, and these are exactly the lines in the shifted cone $x + C_\alpha$. The theorem implies that each of the operators is related to a simple convolution with a singular weighted integration over the cone C_α.

Proof. The theorem is first proven for $f \in \mathcal{D}(\mathbb{R}^3)$, by calculating (30):

$$\begin{aligned}\mathcal{P}_S^*\mathcal{P}_S f(x) &= \int_{\omega \in S}\int_{t \in \mathbb{R}} f(x - (x\cdot\omega)\omega + t\omega)\mathrm{d}t\,\mathrm{d}\omega \\ &= \int_{\omega \in S}\int_{t \in \mathbb{R}} f(x + s\omega)\,\mathrm{d}s\,\mathrm{d}\omega\end{aligned}$$

where the substitution $s = t - (x \cdot \omega)$ is used. Now, one converts this to an integral over C_α:

$$\mathcal{P}_S^* \mathcal{P}_S f(x) = \int_{\omega \in S} \int_{s \in \mathbb{R}} f(x + s\omega) \frac{1}{|s|} |s| \, \mathrm{d}s \mathrm{d}\omega$$
$$= \int_{y \in C_\alpha} \frac{f(x+y)}{\|y\|} \mathrm{d}y$$

since the measure on the cone C_α is $dy = |s| \, ds \, \mathrm{d}\omega$ where $y = s\omega$.

To prove (31), one moves Δ inside the integral. Then one uses rotation invariance of Δ (to write Δ in coordinates:

$$(s, t, p) \mapsto (s\omega + t\omega' + p\omega \times \omega')$$

where ω' is the unit vector tangent to S at ω and in direction of increasing φ). Finally, one uses an integration by parts to show that $\mathcal{P}_S^*$ intertwines Δ and Δ_S. To prove (32), one uses (31) and a calculation to show that $\left(-\Delta + \csc^2(\alpha) \frac{\partial^2}{\partial z^2}\right)$ and $\mathcal{D}_S$ are intertwined by $\mathcal{P}_S^*$.

The theorem is now proven for $f \in \mathcal{E}'(\mathbb{R}^3)$. Since f has compact support, the convolution $f * I$ is defined by [86, 6.37 Theorem]. Then, the proof is completed using the fact that the equalities are true for $f \in \mathcal{D}(\mathbb{R}^3)$ and continuity of the operators (since $\mathcal{P}_S$ is a Fourier integral operator, which will be discussed in section "Microlocal Analysis of Conical Tilt Electron Microscope Tomography (ET)"). □

4 Microlocal Analysis

The reader has seen how limited data reconstructions can image different singularities (or add artifacts) depending on the type of data. This section will be devoted to developing the mathematics to understand those differences. The key point is to develop a precise definition of singularities and to understand what our tomographic transforms do to singularities.

Singular Support and Wavefront Set

Definition 1. Let $\Omega \subset \mathbb{R}^n$ be an open set and let $u \in \mathcal{D}'(\Omega)$. The singular support of u, denoted by $\mathrm{ssupp}(u)$, is the complement in Ω of the largest open set on which u is C^∞ smooth.

In other words, a point $x_0 \in \Omega$ is not in the singular support of u if u is smooth in a neighborhood of x_0. Let us consider some examples.

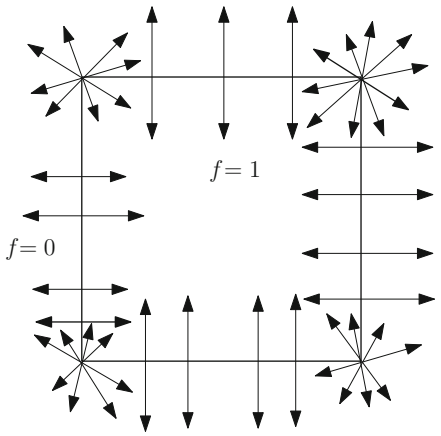


Fig. 9 The function $f = 1$ in the interior of the square and $f = 0$ in the complement. The singular support, $\mathrm{ssupp}(f)$, is the boundary of the rectangle, and the wavefront set directions are shown in the figure

Example 3. Consider the square $S = [0,1]^2$ in $\mathbb{R}^2$. Let f be the characteristic function

$$u(x,y) = \begin{cases} 1 & \text{if } (x,y) \in S; \\ 0 & \text{otherwise.} \end{cases} \tag{33}$$

Then $\mathrm{ssupp}(u)$ is the boundary of the square because that is where u is not smooth; see Fig. 9.

Smoothness of a distribution $u \in \mathcal{E}'(\mathbb{R}^n)$ is related to the rapid decay of its Fourier transform. Recall the following definition.

Definition 2. A function $f : \mathbb{R}^n \to \mathbb{C}$ is *rapidly decaying at infinity* if for every $N \geq 0$, there is a C_N such that $|f(x)| \leq C_N(1 + \|x\|)^{-N}$ for all $x \in \mathbb{R}^n$.

Theorem 11 ([86]). *A distribution $u \in \mathcal{E}'(\Omega)$ is in $C_c^\infty(\Omega)$ if and only if its Fourier transform is rapidly decaying at infinity.*

This theorem implies that if a distribution u is not C^∞ smooth, then there are nonzero frequency directions ξ such that the Fourier transform $\hat{u}$ does not satisfy the estimate of Theorem 11 in any conic neighborhood Γ containing ξ. However, this is global information; it does not yet relate to singular support – the points where u is not smooth. To make this connection, one needs to consider directions near which a localized Fourier transform of u does not satisfy these estimates. This leads us to the concept of C^∞ wavefront set.

Definition 3. Let u be a distribution defined on an open set $\Omega \subset \mathbb{R}^n$. One says that $(x_0, \xi_0) \in \Omega \times \mathbb{R}^n \setminus \mathbf{0}$ is *not* in the wavefront set of u, if there is a $\psi \in C_c^\infty(X)$ identically 1 near x_0 and an open cone Γ containing ξ_0 such that given any N, there is a C_N such that

$$\left|\widehat{\psi u}(\xi)\right| \leq C_N(1 + \|\xi\|)^{-N} \text{ for } \xi \in \Gamma.$$

On the other hand, the C^∞ wavefront set of a distribution u will be denoted by $\mathrm{WF}(u)$.

Remark 1. To be more precise, one views elements of the wavefront set to be elements of the cotangent bundle $T^*\Omega \setminus \mathbf{0}$ using the notation

$$\xi \mathrm{d}x = \xi_1 \mathrm{d}x_1 + \cdots + \xi_n \mathrm{d}x_n \text{ for } \xi \in \mathbb{R}^n \setminus \mathbf{0}.$$

Through this, one can define wavefront sets for distributions on manifolds.

Note that the cutoff function, ψ, in this definition is somewhat more restrictive than what is sometimes given (just that $\psi(x_0) \neq 0$) but it is equivalent [47].

Theorem 12 ([47]). *Let u be a distribution defined on an open set $\Omega \subset \mathbb{R}^n$ and π_x denote the x-projection* $\mathrm{WF}(u)$. *Then* $\pi_x\,(\mathrm{WF}(u)) = \mathrm{ssupp}(u)$.

Example 4. Consider the Dirac delta distribution δ_0 in $\mathbb{R}^n$. Then $\mathrm{ssupp}(\delta_0) = \{0\}$ because δ_0 is zero away from the origin and supported at the origin. So, by Theorem 12, $x = 0$ is the only point above in which there can be wavefront set. Furthermore, if ψ is a cutoff function at $x_0 = 0$, then $\mathcal{F}(\psi\delta_0) = 1/(2\pi)^{n/2}$, so $\mathrm{WF}(\delta_0) = \{(0, \xi \mathrm{d}x), \xi \neq 0\}$.

Example 5. Consider the function f given in Example 3. The wavefront set $\mathrm{WF}(f)$ consists of the nonzero normal directions at all the singular support points except the four corner points. At these corner points, all nonzero directions are in the wavefront set, as illustrated in Fig. 9.

The proof is as follows. Consider first a non-corner point x_0 in the singular support. One can assume that this point is on the x-axis, $x_0 = (a, 0)$, where $a \in (0, 1)$. Fix a direction $\xi^0 = (\xi_1^0, \xi_2^0)$ with $\xi_1^0 \neq 0$. One can show that the localized Fourier transform is rapidly decaying in a conic neighborhood of ξ^0. To see this, one can find a narrow conic neighborhood Γ containing ξ^0 such $|\xi_1| \geq c\,\|\xi\|$ for some $c > 0$ and all $\xi \in \Gamma$ (here $\xi = (\xi_1, \xi_2)$). Let $\varphi \in C_c^\infty(\mathbb{R}^2)$ be a function of the form $\varphi(x_1, x_2) = \varphi_1(x_1)\varphi_2(x_2)$ that is identically 1 near x_0. Without loss of generality, one can assume φ_1 is even about a and φ_2 is even about 0. Consider

$$\widehat{\varphi f}(\xi) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\mathrm{i}x_1 \cdot \xi_1} \varphi_1(x_1) \mathrm{d}x_1 \frac{1}{\sqrt{2\pi}} \int_{0}^{\infty} e^{-\mathrm{i}x_2 \xi_2} \varphi_2(x_2) \mathrm{d}x_2 \,.$$

Denote the left-hand integral in this expression by $L(\xi_1)$ and the right-hand integral by $R(\xi_2)$. Note that the R is bounded in ξ_2 because the integrand is uniformly bounded and of compact support. Because φ_1 is in $\mathcal{S}(\mathbb{R})$, $L(\xi_1)$ is also in $\mathcal{S}(\mathbb{R})$, since it is the one-dimensional Fourier transform of φ_1. Therefore, $L(\xi_1)$ is rapidly decaying at infinity as a function of ξ_1. Since $|\xi_1| > c \|\xi\|$ in Γ, the function $\xi \mapsto L(\xi_1)$ decays rapidly at infinity for ξ in Γ. Since the function $R(\xi_2)$ is bounded, one sees that $\widehat{\varphi f}(\xi)$ decays rapidly in Γ. This shows that the only possible vectors in $\mathrm{WF}(f)$ above $x_0 = (a, 0)$ are vertical ones.

Since f is not smooth at x_0, at least one vertical vector at x_0 must be in $\mathrm{WF}(f)$ by Theorem 11. Therefore, $R(\xi_2)$ must not rapidly decay in either the positive direction ($\xi_2 > 0$) or the negative direction. Since φ_2 is an even function in $\mathcal{S}(\mathbb{R})$, $\mathcal{F}(\varphi_2)(\xi_2) = R(\xi_2) + R(-\xi_2)$ is rapidly decreasing at $\pm\infty$, so $R(\xi_2)$ must *not* be rapidly decaying for $\xi_2 > 0$ *and* for $\xi_2 < 0$ (since $R(\xi_2)$ is not rapidly decaying in at least one direction and the sum is rapidly decaying in both positive and negative directions). Therefore, both vertical vectors are in $\mathrm{WF}(f)$ at x_0. (In another proof, one shows $R(\xi_2) = \mathcal{O}(1/|\xi_2|)$ by performing two integrations by parts on that integral.)

Now it will be shown that all directions are in $\mathrm{WF}(f)$ above $(0, 0)$. One can use symmetric cutoffs in x_1 and x_2 at 0. Then

$$\widehat{\varphi f}(\xi) = \frac{1}{\sqrt{2\pi}} \int_{0}^{\infty} e^{-\mathrm{i}x_2 \xi_2} \varphi_2(x_2) \mathrm{d}x_2 \frac{1}{\sqrt{2\pi}} \int_{0}^{\infty} e^{-\mathrm{i}x_1 \cdot \xi_1} \varphi_1(x_1) \mathrm{d}x_1.$$

The proof for $R(\xi_2)$ above can be used to show that neither integral decays rapidly at infinity in this case. This shows that $\mathrm{WF}(\varphi f)$ consists of all directions at $x_0 = 0$. The proofs at the other corners are similar.

Example 6. If f is the characteristic function of a set, Ω with a smooth boundary, then $\mathrm{WF}(f)$ is the conormal bundle:

$$N^*(\Omega) = \{(x, \xi \mathrm{d}x) \,:\, x \in \mathrm{bd}(\Omega),\ \xi \text{ is normal to } \mathrm{bd}(\Omega) \text{ at } x\}$$

This is suggested by Example 5 and it follows from the results in [47].

If f is a linear combination of characteristic functions of sets with smooth boundary, then $\mathrm{WF}(f)$ is the union of the conormal sets of the individual sets unless cancelation occurs along shared boundaries.

Pseudodifferential Operators

To motivate the definition of these operators, consider the following example.

Theorem 13. *For $f \in \mathcal{S}(\mathbb{R}^2)$,*

$$\mathcal{R}_L^* \mathcal{R}_L u(x) = \int e^{\mathrm{i}x\cdot\xi} \frac{2}{\|\xi\|} \hat{u}(\xi) \mathrm{d}\xi = \frac{1}{\pi} \int e^{\mathrm{i}(x-y)\cdot\xi} \frac{1}{\|\xi\|} u(y) \mathrm{d}y \mathrm{d}\xi \,. \tag{34}$$

Proof. Using a polar integration about x, one shows that

$$\mathcal{R}_L^* \mathcal{R}_L f = f * \frac{2}{\|x\|} \,. \tag{35}$$

Then, since $\mathcal{F}(1/\|x\|) = 1/\|\xi\|$ [43, Lemma 6.2, p. 238], one sees that

$$\mathcal{R}_L^* \mathcal{R}_L u = \mathcal{F}^{-1} \mathcal{F} \left(u * \frac{2}{\|x\|} \right) = \mathcal{F}^{-1} \left(2\pi \frac{2}{\|\xi\|} \hat{u} \right)$$

using the fact that the Fourier transform of a convolution in $\mathbb{R}^2$ (with the normalization of (10)) is the product of the Fourier transforms times 2π. Writing $\mathcal{F}^{-1}$ as an integral in the right-hand expression proves the theorem. □

It should be pointed out that the left-hand integral in (34) converges for $f \in \mathcal{S}(\mathbb{R}^2)$, but the right-hand integral in (34) does not converge. However, one can do integrations by parts at infinity to make it converge for $f \in \mathcal{S}(\mathbb{R}^2)$ or $f \in \mathcal{E}'(\mathbb{R}^2)$ as is done for pseudodifferential operators (e.g., [74]).

With this as the model, consider operators with integral representation

$$\mathcal{P}u(x) = \int e^{\mathrm{i}(x-y)\cdot\xi} p(x, y, \xi) u(y) \mathrm{d}y \mathrm{d}\xi. \tag{36}$$

The study of the operator $\mathcal{P}$ is important in imaging for the following reasons:

1. Assuming p satisfies certain estimates (see Definition 4), one can describe precisely the action of $\mathcal{P}$ on the singularities of u.
2. If one has a procedure to invert or approximately invert the operator $\mathcal{P}$ by another operator $\mathcal{Q}$ having a similar integral representation as that of $\mathcal{P}$, then by (a), one would have that the singularities of $\mathcal{Q}\mathcal{P}u$ are identical to those of u. Now through this approximate inversion process, one has a procedure to recover the singularities of u.

An operator $\mathcal{P}$ of the form (36) with p satisfying certain estimates is called a *pseudodifferential operator* (ΨDO) [33, 47, 89, 93, 94]. This will be given below (see Definition 5).

In order to motivate the appropriate conditions and estimates that p should satisfy, let us look at the following simple example:

Consider a linear partial differential operator of the form

$$\mathcal{P}(x, D) = \sum_{|\nu| \leq m} a_\nu(x) D_x^\nu. \tag{37}$$

Here $\nu = (\nu_1, \cdots, \nu_n)$ is a multi-index and

$$D_x^\nu = (-i)^{(\nu_1 + \cdots + \nu_n)} \frac{\partial^{\nu_1}}{\partial x_1^{\nu_1}} \cdots \frac{\partial^{\nu_n}}{\partial x_n^{\nu_n}}.$$

For simplicity let u be a compactly supported function. Applying the Fourier transform,

$$\widehat{D_x^\nu u}(\xi) = \frac{1}{(2\pi)^{n/2}} \int e^{-ix\cdot\xi} D_x^\nu u(x) \mathrm{d}x. \tag{38}$$

Integrating by parts $|\nu|$ times, one obtains

$$\widehat{D_x^\nu u}(\xi) = \xi^\nu \hat{u}(\xi). \tag{39}$$

With this one has

$$\mathcal{P}(x, D)u(x) = \frac{1}{(2\pi)^{n/2}} \int e^{\mathrm{i}(x-y)\cdot\xi} p(x, \xi) u(y) \mathrm{d}y \mathrm{d}\xi \tag{40}$$

where

$$p(x, \xi) = \sum_\nu a_\nu(x) \xi^\nu.$$

The function $p(x, \xi)$ is called the *symbol* of the partial differential operator (PDO), $\mathcal{P}(x, D)$.

The function $p(x, \xi)$ satisfies the following property: Let $\alpha = (\alpha_1, \cdots, \alpha_n)$ be a multi-index. Then $\partial_\xi^\alpha p(x, \xi)$ lowers the degree with respect to ξ of the resulting function by $|\alpha|$, whereas $\partial_x^\beta p(x, \xi)$ where $\beta = (\beta_1, \cdots, \beta_n)$ is a multi-index does not alter the degree of homogeneity with respect to ξ of the resulting function.

More precisely one has the following estimate:

Let α and β be any multi-indices. For x in a bounded subset of $\mathbb{R}^n$, there is a constant C such that

$$\left|\partial_\xi^\alpha \partial_x^\beta p(x,\xi)\right| \leq C(1+\|\xi\|)^{m-|\alpha|}. \tag{41}$$

Here m is the order of the PDO, $\mathcal{P}(x, D)$. In order to get this inequality, first rewrite the terms of $p(x,\xi)$ by combining terms of the same homogeneous degree with respect to the ξ variable.

Differentiate ξ^ν $\alpha = (\alpha_1, \cdots, \alpha_n)$ times with respect to $\xi_1, \cdots, \xi_n$, where the number of times one differentiates ξ^ν with respect to each particular ξ_l is α_l. One sees that this reduces the degree of homogeneity of ξ^ν by $|\alpha|$ and the highest-order terms dominate. On a bounded subset of $\mathbb{R}^n$, all derivatives of the a_ν are bounded. This gives the estimate (41).

Now one can generalize the class of operators that have Fourier integral representations of the form (40) by admitting a larger class of functions $p(x,\xi)$ to be symbols. Consider those functions p that satisfy the estimate as in (41) and that behave like polynomials or the inverse of polynomials in ξ as $\|\xi\| \to \infty$. In other words, one wants $p(x,\xi)$ to grow or decay in powers of $\|\xi\|$, and differentiation with respect to ξ lowers the order of growth or raises the order of decay. Furthermore, in order to include $\mathcal{R}_L^*\mathcal{R}_L$ in the class of operators considered (see Example 13), one can allow some latitude at $\xi = 0$.

In the interest of flexibility, one can also let the function p depend on x, y, and ξ. Such functions will be denoted as *amplitudes*.

Definition 4 ([33, 47, 89, 93, 94]). Let $X \subset \mathbb{R}^n$ be an open subset. An amplitude of order m is a function that satisfies the following properties:

1. $p(x,y,\xi) \in C^\infty(X \times X \times \mathbb{R}^n \setminus \{0\})$,
2. For every compact set $K \subset X$ and for multi-index α, β, γ,
   (a) There is a constant $C = C(K, \alpha, \beta, \gamma)$ such that

$$\left|D_\xi^\alpha D_x^\beta D_y^\gamma p(x,y,\xi)\right| \leq C(1+\|\xi\|)^{m-|\alpha|} \text{ for } x \text{ and } y \text{ in } K \text{ and } \|\xi\| > 1, \text{ and}$$

   (b) $p(x,y,\xi)$ is locally integrable for x and y in K and $\|\xi\| \leq 1$.

It is important to note that in Definition 4, p need not be a polynomial in ξ and m can be any real number. The local integrability condition can be relaxed if p is a sum of homogeneous terms in ξ [74].

Now pseudodifferential operators are defined.

Definition 5. Let $X \subset \mathbb{R}^n$ be an open subset. A pseudodifferential operator (ΨDO) is an operator of the form

$$\mathcal{P}u(x) = \frac{1}{(2\pi)^n} \int e^{\mathrm{i}(x-y)\cdot\xi} p(x,y,\xi) u(x) \mathrm{d}y \mathrm{d}\xi,$$

where $p(x,y,\xi)$ is a function that satisfies the properties of Definition 4.

The operator $\mathcal{P}$ has order m if its amplitude is of order m, and $\mathcal{P}$ is elliptic of order m if for each compact set $K \subset \Omega$, there is a constant $C_K > 0$ such that for x and y in K and $\|\xi\| \geq C_K$

$$|p(x, y, \xi)| \geq C_K^{-1}(1 + \|\xi\|)^m . \tag{42}$$

The next theorem highlights two fundamental properties of ΨDOs.

Theorem 14 (Pseudolocal Property [94]). *If $\mathcal{P}$ is a ΨDO, then $\mathcal{P}$ satisfies the* pseudolocal property*:*

$$\operatorname{ssupp}(\mathcal{P}u) \subset \operatorname{ssupp}(u) \text{ and } \operatorname{WF}(\mathcal{P}u) \subset \operatorname{WF}(u) .$$

If, in addition, $\mathcal{P}$ is elliptic, then

$$\operatorname{ssupp}(\mathcal{P}u) = \operatorname{ssupp}(u) \text{ and } \operatorname{WF}(\mathcal{P}u) = \operatorname{WF}(u) .$$

Note that, although ΨDOs can spread out the support of the function u, they do not spread out its singular support or wavefront set. Elliptic ΨDOs preserve singular support and wavefront set.

Equation (34) shows that the composition $\mathcal{R}_L^* \mathcal{R}_L$ is a pseudodifferential operator since its symbol $(4\pi)/\|\xi\|$ satisfies the conditions in Definition 4. Furthermore, because the symbol satisfies the ellipticity estimate (42), $\mathcal{R}_L^* \mathcal{R}_L$ is an elliptic ΨDO of order -1.

Example 7. The powers of d/dp and Δ as ΨDOs.

According to (40), the symbol of $-d^2/dp^2$ is $|\tau|^2$ where τ is the dual variable to p. So, the symbol of $\sqrt{-d^2/dp^2}$ can be considered to be $|\tau|$. The rationale is that if one calculates

$$\begin{aligned} -\frac{d^2}{dp^2} f &= \mathcal{F}^{-1} |\tau|^2 \mathcal{F} f = \mathcal{F}^{-1} |\tau| \mathcal{F} \left(\mathcal{F}^{-1} |\tau| \mathcal{F} f \right) \\ &= \sqrt{-\frac{d^2}{dp^2}} \circ \sqrt{-\frac{d^2}{dp^2}} f . \end{aligned} \tag{43}$$

This justifies why

$$\Lambda_p = \sqrt{-d^2/dp^2} . \tag{44}$$

Since

$$-\Delta = D_{x_1}^2 + \cdots + D_{x_n}^2 ,$$

its symbol is $\|\xi\|^2$. Since the symbol of $-\Delta$ is $\|\xi\|^2$, one can easily find symbols of powers of the Laplacian. For example, $\sqrt{-\Delta}$ has symbol $\|\xi\|$ and Fourier representation

$$\sqrt{-\Delta}u = \frac{1}{(2\pi)^n}\int e^{\mathrm{i}(x-y)\cdot\xi}\,\|\xi\|\,u(y)\mathrm{d}y\mathrm{d}\xi$$

and $(-\Delta)^{-1/2}$ has symbol $1/\|\xi\|$

$$(-\Delta)^{-1/2}u = \frac{1}{(2\pi)^n}\int e^{\mathrm{i}(x-y)\cdot\xi}\frac{1}{\|\xi\|}u(y)\mathrm{d}y\mathrm{d}\xi.$$

Now, as a calculation in distributions using Fourier transforms (similar to (43)), one sees that $(-\Delta)^{-1/2}\circ\sqrt{-\Delta}$ is the identity map.

Note that $\sqrt{-d^2/dp^2}$, $\sqrt{-\Delta}$, $(-\Delta)^{-1/2}$ are all elliptic ΨDOs since the corresponding symbols satisfy the ellipticity estimate (42).

Example 8. Example 7 allows one to justify why the operator Λ_p is really $\sqrt{-d^2/dp^2}$. So, the FBP inversion formula (18) can be written

$$f = \frac{1}{4\pi}\mathcal{R}_L^*\left(\sqrt{-d^2/dp^2}\mathcal{R}_L f\right). \tag{45}$$

Equation (34) shows that $\mathcal{R}_L^*\mathcal{R}_L = 4\pi\,(-\Delta)^{-1/2}$. Using the observation at the end of Example 7, one obtains a different version of the filtered backprojection inversion formula for $\mathcal{R}_L$:

$$f = \frac{1}{4\pi}\sqrt{-\Delta}\mathcal{R}_L^*\mathcal{R}_L f\,. \tag{46}$$

These calculations can be justified for distributions of compact support [43, 45].

Example 9. Now consider the Lambda operators given in (22) and (23). To get the Lambda operator, $\mathcal{L}_x$, from the FBP operator, one replaces the $\sqrt{-d^2/dp^2}$ in (45) by its square, $-d^2/dp^2$.

Here is another way to understand Lambda tomography. By evaluating another $\sqrt{-\Delta}$ in (46), one sees that

$$\sqrt{-\Delta}f = -\frac{1}{4\pi}\Delta\mathcal{R}_L^*\mathcal{R}_L f = \frac{1}{4\pi}\mathcal{R}_L^*\left(-\frac{d^2}{dp^2}\mathcal{R}_L f\right)$$

where the second equality holds because $\mathcal{R}_L^*$ intertwines $-\Delta$ and $-d^2/dp^2$ (this is proven using an argument similar to the intertwining argument in the proof of Theorem 10). Because $\mathcal{R}_L^*\mathcal{R}_L$ is an elliptic ΨDO with symbol $4\pi/\|\xi\|$, the symbol of $\mathcal{L}_{x,\mu}$ is

$$\|\xi\| + \frac{\mu}{\|\xi\|}$$

and it is elliptic of order one. Therefore, $\mathcal{L}_{x,\mu}$ and $\mathcal{L}_x$ (corresponding to $\mu = 0$) are both elliptic ΨDOs.

The operator $\mathcal{L}_{x,\mu}$ does not reconstruct f but $\left((-\Delta)^{1/2} + \mu\right) f$. The natural question then is how different is this from f? It has just been established that $\mathcal{L}_x$ and $\mathcal{L}_{x,\mu}$ are elliptic ΨDOs. Therefore, by Theorem 14 this means that these operators recover $\mathrm{ssupp}(f)$ and $\mathrm{WF}(f)$. The reconstructions in Fig. 8 and others in [17, 18] show that it is effective in practice.

Fourier Integral Operators

In the analysis thus far, one studied the composition of a generalized Radon transform with its adjoint. This composed operator, as was shown, is a pseudodifferential operator. Theorem 14 shows us how ΨDOs act on singularities and wavefront sets. In this section, more general operators will be considered, and how they change wavefront sets will be described.

Example 10. Consider $\mathcal{R}_L$ in a special Fourier representation.

$$\begin{aligned} \mathcal{R}_L f(\omega, p) &= \frac{1}{(2\pi)^{1/2}} \int_{\tau\in\mathbb{R}} e^{\mathrm{i}p\tau} \mathcal{F}_p \left(\mathcal{R}_L f\right)(\omega, \tau)\, \mathrm{d}\tau \\ &= \int_{\tau\in\mathbb{R}} e^{\mathrm{i}p\tau} \hat{f}(\tau\omega)\mathrm{d}\tau \\ &= \int_{\tau\in\mathbb{R}} \int_{x\in\mathbb{R}^2} e^{\mathrm{i}(p-(x\cdot\omega))\tau} \frac{1}{2\pi} f(x)\mathrm{d}x\mathrm{d}\tau\,. \end{aligned} \tag{47}$$

The last expression in (47) looks like a ΨDO except that the τ and x integrals are over different sets and the exponent is not the one for ΨDOs.

In many applications, it might be necessary to understand the properties of the Radon transform directly, rather than the composition with its adjoint. In the last example one saw that the Radon transform $\mathcal{R}_L$ had an integral representation of the form

$$\mathcal{P}u(x) = \int e^{\mathrm{i}\phi(x,y,\xi)} p(x, y, \xi) u(y)\mathrm{d}y\mathrm{d}\xi. \tag{48}$$

The important differences between the operator $\mathcal{P}$ in (48) and a ΨDO are the following:

- The functions $\mathcal{P}u$ and u, in general, are functions on different sets X and Y, respectively. The spaces X and Y can be of different dimensions as well.
- The phase function ϕ in (48) is more general than that of a ΨDO, but it shares similar features. See Definition 6.
- The dimension of the frequency variable ξ can be different from that of the spaces X and Y, unlike as in the case of a ΨDO.

Another simple example where an integral representation of (48) arises is when one uses Fourier transform techniques to determine the solution to a constant coefficient wave equation:

$$\left(\partial_t^2 - \Delta_x\right) u = 0, \quad u(x,0) = 0, \quad \partial_t u(x,0) = g. \tag{49}$$

Now by taking Fourier transform in the space variable, one has the following integral representation for the solution to the wave equation:

$$u(x,t) = \frac{1}{2\mathrm{i}(2\pi)^n}\Big(\int e^{\mathrm{i}(x-y)\cdot\xi + t\|\xi\|} \frac{1}{\|\xi\|} g(y)\mathrm{d}y\mathrm{d}\xi - \int e^{\mathrm{i}(x-y)\cdot\xi - t\|\xi\|} \frac{1}{\|\xi\|} g(y)\mathrm{d}y\mathrm{d}\xi \Big).$$

Note that the phase functions in the above solution are $\phi_\pm(x,y,\xi) = (x-y)\cdot\xi \pm t\,\|\xi\|$.

If $f : \mathbb{R}^n \to \mathbb{R}$, then the notation

$$\partial_x f(x) = \frac{\partial f}{\partial x_1}\mathrm{d}x_1 + \cdots + \frac{\partial f}{\partial x_1}\mathrm{d}x_n \tag{50}$$

will be used for the differential of f with respect to x. If g is a function of (φ, p), then

$$\partial_{\varphi,p} g(\varphi,p) = \frac{\partial g}{\partial \varphi}\mathrm{d}\varphi + \frac{\partial g}{\partial p}\mathrm{d}p$$

will denote the differential of g with respect to the variables (φ, p).

Definition 6. Let $X \subset \mathbb{R}^m$ and $Y \subset \mathbb{R}^n$ be open subsets. A real-valued function $\phi \in C^\infty(X \times Y \times \mathbb{R}^N \setminus \{0\})$ is called a phase function if

1. ϕ is positive homogeneous of degree 1 in ξ. That is, $\phi(x,y,r\xi) = r\phi(x,y,\xi)$ for all $r > 0$.
2. $(\partial_x\phi, \partial_\xi\phi)$ and $(\partial_y\phi, \partial_\xi\phi)$ do not vanish for all $(x,y,\xi) \in X \times Y \times \mathbb{R}^N \setminus \{0\}$.

The phase function ϕ is called nondegenerate if on the zero-set $\partial_\xi \phi = 0$, the set of vectors $\{\partial_{x,y,\xi} \frac{\partial \phi}{\partial \xi_j}, 1 \leq j \leq N\}$ is linearly independent.

Definition 7. A Fourier integral operator (FIO) $\mathcal{P}$ is defined as

$$\mathcal{P}u(x) = \int e^{i\phi(x,y,\xi)} p(x,y,\xi) u(y) \mathrm{d}y \mathrm{d}\xi,$$

where ϕ is a nondegenerate phase function, and the amplitude $p(x,y,\xi) \in C^\infty(X \times Y \times \mathbb{R}^N)$ satisfies the following estimate: for every compact set $K \subset X \times Y$ and for every multi-index α, β, γ, there is a constant $C = C(K, \alpha, \beta, \gamma)$ such that

$$\left| D_\xi^\alpha D_x^\beta D_y^\gamma p(x,y,\xi) \right| \leq C(1 + \|\xi\|)^{m-|\alpha|} \text{ for all } x, y \in K \text{ and for all } \xi \in \mathbb{R}^N.$$

Finally, one has two important sets associated with this FIO.

The auxiliary manifold

$$\Sigma_\phi = \left\{(x,y,\xi) \in X \times Y \times \left(\mathbb{R}^N \setminus \mathbf{0}\right) : \partial_\xi \phi(x,y,\xi) = 0\right\} \tag{51}$$

and the *canonical relation*

$$C := \left\{\left(x, \partial_x \phi(x,y,\xi); y, -\partial_y \phi(x,y,\xi)\right) : (x,y,\xi) \in \Sigma_\phi\right\}. \tag{52}$$

Note that since the phase function ϕ is nondegenerate, the sets Σ_ϕ and C have smooth manifold structures.

One can also include the local integrability condition 2b of Definition 4 for the amplitude of FIOs.

Note that $\mathcal{R}_L$ satisfies these conditions with phase function $\phi(\omega, p, x, \tau) = (p - x \cdot \omega)\tau$ and amplitude $p(x,y,\tau) = 1/(2\pi)$, so $\mathcal{R}_L$ is an FIO. Guillemin originally proved that a broad range of Radon transforms are FIOs [35, 36, 38]. The operator $\mathcal{R}_L$ will be studied more carefully in the next section.

Every ΨDO is an FIO with phase function $\phi(x,y,\xi) = (x - y) \cdot \xi$. However, $\mathcal{R}_L$ is an FIO that is not a ΨDO since its phase function is not of that form.

Definition 8. Let $C \subset T^*X \times T^*Y$, $\tilde{C} \subset T^*Y \times T^*X$, and $A \subset T^*Y$. Define

$$C \circ A = \left\{(x, \xi \mathrm{d}x) : \exists\, (y, \eta \mathrm{d}y) \in A \text{ with } (x, \xi \mathrm{d}x; y, \eta \mathrm{d}y) \in C\right\}$$

$$\tilde{C} \circ C = \Big\{(\tilde{y}, \tilde{\eta} \mathrm{d}\tilde{y}; y, \eta \mathrm{d}y) : \exists\, (x, \xi \mathrm{d}x) \text{ with } (\tilde{y}, \tilde{\eta} \mathrm{d}\tilde{y}; x, \xi \mathrm{d}x) \in \tilde{C},$$
$$\text{and } (x, \xi \mathrm{d}x; y, \eta \mathrm{d}y) \in C\Big\}.$$

The key theorem on what FIOs do to wavefront sets is the following result.

Theorem 15 ([47]). *Let $\mathcal{P}$ be an FIO and let C be the associated canonical relation. Then*

$$\mathrm{WF}(\mathcal{P}u) \subset C \circ \mathrm{WF}(u).$$

Example 11. In this example, the canonical relation of any ΨDO will be calculated. First note that since $\phi = (x - y) \cdot \xi$,

$$\partial_x \phi = \xi \mathrm{d}x \quad \partial_y \phi = -\xi \mathrm{d}y \quad \partial_\xi \phi = (x - y)\mathrm{d}\xi.$$

Therefore, $\Sigma_\phi = \{(x, y, \xi) : x - y = 0\}$. Now, use (52) and these calculations to see that the canonical relation for ΨDOs is

$$C = \{(x, \xi \mathrm{d}x; x, \xi \mathrm{d}x) : \xi \neq 0\}$$

which is the diagonal in $(T^*(\mathbb{R}^n) \setminus \mathbf{0})$ and which will be denoted by Δ. Using Theorem 15, one sees that if $\mathcal{P}$ is a ΨDO, $\mathrm{WF}(\mathcal{P}u) \subset \mathrm{WF}(u)$.

This section will be concluded with a key theorem on the wavefront set of the composition of two FIOs known as the Hörmander-Sato Lemma.

Theorem 16 (Hörmander-Sato Lemma [47]). *Let $\mathcal{P}_1$ and $\mathcal{P}_2$ be two FIOs with canonical relations C_1 and C_2, respectively. Assume $\mathcal{P}_1 \circ \mathcal{P}_2$ is defined for distributions of compact support, and let u be a distribution of compact support. Then*

$$\mathrm{WF}((\mathcal{P}_1 \circ \mathcal{P}_2)u) \subset (C_1 \circ C_2) \circ \mathrm{WF}(u).$$

From an imaging point of view, the operator that is studied is the image reconstruction operator $\mathcal{P}^*\mathcal{P}$ where $\mathcal{P}^*$ is the adjoint of the FIO $\mathcal{P}$. Using Hörmander-Sato Lemma, one can study the wavefront set of the image reconstruction operator. For this one requires the canonical relation of the adjoint $\mathcal{P}^*$ [46] which is given by

$$C^t = \{(y, \eta \mathrm{d}y; x, \xi \mathrm{d}x) : (x, \xi \mathrm{d}x; y, \eta \mathrm{d}y) \in C\}.$$

Now from Hörmander-Sato Lemma, when the composition is defined, one has

$$\mathrm{WF}(\mathcal{P}^*\mathcal{P}f) \subset \left(C^t \circ C\right) \circ \mathrm{WF}(f).$$

5 Applications to Tomography

In this section microlocal analysis will be used to explain strengths and limitations of the reconstructions from Sect. 2.

Microlocal Analysis in X-Ray CT

As seen in section "X-Ray Tomography (CT) and Limited Data Problems," reconstructions from different limited data problems have different strengths and weaknesses. The information in the last section will now be used to understand the microlocal properties of $\mathcal{R}_L$.

To make the cotangent space calculations simpler, the following coordinates will be used on $S^1 \times \mathbb{R}$: $(\varphi, p) \mapsto (\omega(\varphi), p)$ where $\omega(\varphi) = (\cos(\varphi), \sin(\varphi))$ and $\omega^\perp(\varphi) = \omega(\varphi + \pi/2)$. Thus, functions on $S^1 \times \mathbb{R}$ will now be written in terms of (φ, p).

Theorem 17. *The Radon transform $\mathcal{R}_L$ is an elliptic FIO associated with the canonical relation*

$$C_L = \Big\{ \big(\varphi, p, \alpha(-x \cdot \omega^\perp(\varphi)\mathrm{d}\varphi + \mathrm{d}p); x, \alpha\omega(\varphi)\mathrm{d}x\big) : (\varphi, p) \in [0, 2\pi] \times \mathbb{R}, x \in \mathbb{R}^2, \alpha \neq 0, x \cdot \omega(\varphi) = p \Big\}. \tag{53}$$

For $f \in \mathcal{E}'(\mathbb{R}^2)$,

$$\mathrm{WF}(\mathcal{R}_L f) = C_L \circ \mathrm{WF}(f)\,. \tag{54}$$

Furthermore, $C_L^t \circ C_L = \Delta$ is the diagonal in $\big(T^*(\mathbb{R}^2) \setminus \mathbf{0}\big)^2$.

Proof. In Example 10, it was shown that $\mathcal{R}_L$ is an FIO associated to the phase function

$$\phi(\varphi, p, x, \tau) = \tau(p - x \cdot \omega(\varphi))\,.$$

To calculate the canonical relation for $\mathcal{R}_L$, the general methods outlined in section "Fourier Integral Operators" (see also [47, p. 165] or [95, (6.1) p. 462]) will be followed. First, the differentials of ϕ will be calculated:

$$\begin{gathered} \partial_x \phi = -\tau\omega(\varphi)\mathrm{d}x, \quad \partial_{(\varphi,p)}\phi = \tau\left(-x \cdot \omega^\perp(\varphi)\mathrm{d}\varphi + \mathrm{d}p\right) \\ \partial_\tau \phi = (p - x \cdot \omega(\varphi))\mathrm{d}\tau. \end{gathered} \tag{55}$$

Note that the conditions for ϕ to be a nondegenerate phase function [95, (2.2)–(2.4), p. 315] hold because $\partial_x\phi$ and $\partial_{(\varphi,p)}\phi$ are not zero for $\tau \neq 0$. Therefore, $\mathcal{R}_L$ is a Fourier integral operator. $\mathcal{R}_L$ has order $-1/2$ because its symbol $1/2\pi$ is homogeneous of degree zero, $2 = \dim \mathbb{R}^2 = \dim Y$, and σ is one dimensional (see [95, p. 462 under (6.3)]). Since the symbol, $1/2\pi$, is homogeneous and nowhere zero, $\mathcal{R}_L$ is elliptic (see [46]).

The auxiliary manifold Σ_ϕ is

$$\Sigma_\phi = \{(\varphi, p, x, \tau) \in ([0, 2\pi] \times \mathbb{R}) \times \mathbb{R}^2 \times (\mathbb{R} \setminus \mathbf{0}) \,:\, p - x \cdot \omega(\varphi) = 0\}. \tag{56}$$

The canonical relation, C_L, associated to $\mathcal{R}_L$ is defined by the map

$$\Sigma_\phi \ni (x, \varphi, p, \tau) \mapsto \left(\varphi, (x \cdot \omega(\varphi)), \partial_{(\varphi,p)}\phi; x, -\partial_x\phi\right).$$

One uses this and a calculation to justify the expression (53).

To show $C_L^t \circ C_L = \Delta$, one starts with $(x, \xi \mathrm{d}x) \in T^*(\mathbb{R}^2)$ and then follows through the calculation of $C_L^t \circ C_L$ using (53): First, choose $\varphi \in [0, 2\pi]$ such that $\xi = a\omega(\varphi)$ for some $a > 0$. Then, there are two vectors associated to $(x, \xi \mathrm{d}x)$ in C_L,

$$\lambda_1 = \left(\varphi, x \cdot \omega(\varphi); a(-x \cdot \omega^\perp(\varphi)\mathrm{d}\varphi + \mathrm{d}p)\right),$$
$$\lambda_2 = \left(\varphi + \pi, x \cdot \omega(\varphi + \pi); -a(-x \cdot \omega^\perp(\varphi + \pi)\mathrm{d}\varphi + \mathrm{d}p)\right).$$

Under C_L^t, λ_1 is associated with $\left((x \cdot \omega(\varphi))\omega(\varphi) + x \cdot \omega^\perp(\varphi), a\omega(\varphi)\mathrm{d}x\right)$ and this is exactly $(x, \xi \mathrm{d}x)$, and there is no other vector in $T^*(\mathbb{R}^2)$ associated with λ_1. In a similar way, one shows that the only vector in $T^*(\mathbb{R}^2)$ associated with λ_2 is $(x, \xi \mathrm{d}x)$. Therefore,

$$C_L^t \circ C_L = \left\{(x, \xi \mathrm{d}x; x, \xi \mathrm{d}x) \,:\, (x, \xi \mathrm{d}x) \in T^*(\mathbb{R}^2) \setminus \mathbf{0}\right\} = \Delta.$$

Because the symbol of $\mathcal{R}_L$, $1/2\pi$, is nowhere zero and homogeneous, $\mathcal{R}_L$ is elliptic [95]. Furthermore, because C_L is a local canonical graph (e.g., [75] and can be seen from the calculation in the last paragraph), a stronger version of the Hörmander-Sato Lemma for elliptic operators [95] can be used to show $\mathrm{WF}(\mathcal{R}_L f) = C_L \circ \mathrm{WF}(f)$. □

Note that the fact $C_L^t \circ C_L = \Delta$ implies that $\mathrm{WF}\left(\mathcal{R}_L^* \mathcal{R}_L(f)\right) \subset \mathrm{WF}(f)$, by the Hörmander-Sato Lemma (Theorem 16). This and the composition calculus of FIO [46, Theorem 4.2.2] provide another proof that $\mathcal{R}_L^* \mathcal{R}_L$ is a ΨDO.

This theorem has the following important corollaries.

Corollary 1 (Propagation of Singularities for $\mathcal{R}_L$). *Let $f \in \mathcal{E}'(\mathbb{R}^2)$.*

1. *Let* $(x_0, \xi_0 \mathrm{d}x) \in T^*(\mathbb{R}^2) \setminus \mathbf{0}$ *and let* φ_0 *be chosen so that* $\xi_0 = \alpha\omega(\varphi_0)$ *for some* $\alpha \neq 0$. *If* $(x_0, \xi_0 \mathrm{d}x) \in \mathrm{WF}(f)$, *then*

$$\left(\varphi_0, x_0 \cdot \omega(\varphi_0); \alpha(-x_0 \cdot \omega^{\perp}(\varphi_0)\mathrm{d}\varphi + \mathrm{d}p)\right) \in \mathrm{WF}(\mathcal{R}_L f).$$

2. *Let* $(\varphi_0, p_0) \in [0, 2\pi] \times \mathbb{R}$ *and* $A \in \mathbb{R}$. *Assume* $(\varphi_0, p_0; \alpha(-A\mathrm{d}\varphi + \mathrm{d}p)) \in \mathrm{WF}(\mathcal{R}_L f)$. *Then,* $(x_0, \xi_0 \mathrm{d}x) \in \mathrm{WF}(f)$ *where* $x_0 = p_0\omega(\varphi_0) + A\omega^{\perp}(\varphi_0)$ *and* $\xi_0 = \alpha\omega(\varphi_0)$.

This provides the paradigm:
$\mathcal{R}_{\mathrm{L}}$ ***detects singularities of*** f ***perpendicular to the line of integration ("visible" singularities), but singularities of*** f ***in other ("invisible") directions do not create singularities of*** $\mathcal{R}_{\mathrm{L}}\mathrm{f}$ ***near the line of integration.***

Remark 2. The paradigm has implications for limited data tomography. A wavefront direction $(x_0, \xi_0 \mathrm{d}x) \in \mathrm{WF}(f)$ will be visible from limited Radon data if and only if the line through x_0 perpendicular to ξ_0 is in the data set.

Proof (Proof of Corollary 1). Using (54), one sees that

$$(x_0, \xi_0 \mathrm{d}x) \in \mathrm{WF}(f) \text{ if and only if } C_L \circ \{(x_0, \xi_0 \mathrm{d}x)\} \in \mathrm{WF}(\mathcal{R}_L f).$$

Part (a) follows from the forward implication of this equivalence and part (b) follows from the reverse implication using the expression for C_L, (53).

Part (a) implies that $\mathcal{R}_L$ detects singularities perpendicular to the line being integrated over, since $\omega(\varphi_0)$ is perpendicular to the line $L(\omega(\varphi_0), x_0 \cdot \omega(\varphi_0))$. Part (b) implies that if a singularity is visible in $\mathcal{R}_L f$ at (φ_0, p_0), it must come from a point on $L(\omega(\varphi_0), p_0)$ and in a direction perpendicular to this line. This explains the paradigm in the theorem. □

Corollary 2 (Propagation of Singularities for Reconstruction Operators). *Let* $f \in \mathcal{E}'(\mathbb{R}^2)$.

(a) Let $\mathcal{L}$ *be either the FBP (see* (18)*), Lambda (*$\mathcal{L}_x$, (22)*), or Lambda + contour (*$\mathcal{L}_{x,\mu}$, (23)*) operator. Let* $f \in \mathcal{E}'(\mathbb{R}^2)$. *Then,* $\mathrm{WF}(f) = \mathrm{WF}(\mathcal{L} f)$.
(b) Let $\mathcal{R}^*_{L,\mathrm{lim}}$ *be the limited angle backprojection operator in* (25) *for angles* $a < \varphi < b$ *(where* $b - a < \pi$*). Let*

$$\mathcal{V} = \{(x, \xi \mathrm{d}x) \,:\, \xi = \alpha\omega(\varphi), \varphi \in (a, b)\}\ .$$

If $\mathcal{L}_{\mathrm{lim}}$ *is any of the operators*

$$\mathcal{R}^*_{L,\mathrm{lim}} \Lambda_p \mathcal{R}_L, \quad \mathcal{R}^*_{L,\mathrm{lim}} \left(-\frac{d^2}{dp^2}\mathcal{R}_L\right), \quad \mathcal{R}^*_{L,\mathrm{lim}} \left(\left(-\frac{d^2}{dp^2} + \mu\right)\mathcal{R}_L\right),$$

then

$$\mathrm{WF}(\mathcal{L}_{\lim} f) \cap \mathcal{V} = \mathrm{WF}(f) \cap \mathcal{V}\,.$$

Note that $\mathcal{V}$ is the set of visible singularities for the limited angle transform.

Proof. Part (a) follows from the fact that both $\mathcal{L}_x$ and $\mathcal{L}_{x,\mu}$ are elliptic, as noted in Example 9 and the equalities stated in Theorem 14.

Part (b) follows from the fact that when one cuts off angles, one can see only wavefront parallel to the angles in the data set, that is, the visible directions in $\mathcal{V}$. A complete proof of this result is given in [26]. □

Remark 3. The paradigms in Corollaries 1 and 2 have especially simple interpretations if f is the sum of characteristic functions of sets with smooth boundaries. The tangent line to any point on the boundary of a region is normal to the wavefront direction of f at that point (since the wavefront set at that point is normal to the boundary, as noted in Example 6).

So, a boundary at x (with conormal $(x, \xi \mathrm{d}x)$) will be visible from the data $\mathcal{R}_L f$ near $(\omega(\varphi), p)$ if (x, ξ) is normal to the line of integration. This can be stated more simply as follows: *the boundary at x is tangent to a line in the data set.*

Finally, note that Example 1 provides a simple case for which $\mathcal{R}_L f$ is not smooth when the line $L(\varphi, p)$ is tangent to the boundary of $\mathrm{supp}(f)$. The paradigm shows this principle is true generally.

Limited Data X-Ray CT

Each of the limited data problems discussed in Sect. 2 will now be examined in light of the corollaries and paradigm of the last section.

Exterior X-Ray CT Data

In the reconstruction in Fig. 2, the boundaries tangent to lines in the data set are clearer and less fuzzy than the one not tangent to lines in the data set. The paradigm in Corollary 1 and Remarks 2 and 3 explains this perfectly. When a line in the data set is tangent to a boundary, then the boundary is visible in the reconstruction. For the exterior reconstruction in Fig. 2, this is true for the inside boundary of the disk at about eight o'clock on the circle (lower left), even though that boundary is imaged by only a few lines in the exterior data set. If the line tangent to the boundary at x is not in the data set, then the boundary is fuzzier, as can be seen in that figure.

This is reflected in Quinto's algorithm [77] in the following way. That algorithm expands the reconstruction in a polar Fourier series

$$f(r\omega(\varphi)) = \sum_{\ell \in \mathbb{Z}} f_\ell(r) e^{i\ell\varphi},$$

where $f_\ell(r)$ is approximated by a polynomial which is calculated using quadrature. The calculation of f_ℓ could be done stably only up to about $|\ell| = 25$. However, for each ℓ for which this could be done, the recovery of $f_\ell(r)$ is very accurate and could be done up to a polynomial of order about 100. Thus, the algorithm has good radial resolution but bad resolution in the polar direction. However, this is, at least in part, a limitation of the problem, not just the algorithm. The singularities in Fig. 2 that are smoothed by the algorithm are intrinsically difficult to reconstruct.

Limited Angle Data

In Fig. 3, data are given over lines $L(\omega(\varphi), p)$ for $\varphi \in [-\pi/4, \pi/4]$, and the only boundaries that are visible are exactly those normal to such lines. This reflects Corollary 1 and Remarks 2 and 3. The algorithm used is FBP but with a limited angle backprojection (see (25) and the discussion below it) between $-\pi/4$ and $\Phi = \pi/4$. In this case, Corollary 2 implies that the only singularities of f that will be visible in $\mathcal{L}_{\text{lim}} f$ (where $\mathcal{L}_{\text{lim}}$ is given in that corollary) are those in the cone $\mathcal{V} = \{r\omega(\varphi) : \varphi \in (-\pi/4, \pi/4), r \neq 0\}$.

However, there is also a marked streak along the lines with angle $\varphi = \pm\pi/4$ that are tangent to the boundary of the region. Frikel and Quinto have recently explained this using microlocal analysis [26], see also [49]. They also explained that in order to decrease those streaks, one needs to make the backprojection operator a standard FIO by including a smooth cutoff function ψ supported in $(-\pi/4, \pi/4)$ that is equal to 1 on most of that interval:

$$\mathcal{R}_{L\psi}^* g(x) = \int_{\varphi=-\pi/4}^{\pi/4} g(\omega(\varphi), x \cdot \omega(\varphi)) \psi(\varphi) \mathrm{d}\varphi .$$

Region of Interest (ROI) Data

The reconstruction in Fig. 8 is from Lambda CT, and all singularities of the circle are visible. This is expected because of Corollary 2 Part (a), the Lambda operator preserves wavefront set. Recall that $\mathcal{L}_x(f)(x)$ determined by data $\mathcal{R}_L f$ for lines near x. This means that the wavefront of f above x (which is the same as the wavefront set of $\mathcal{L}_x(f)$ above x), can be determined by local data for lines near x. This is why all singularities are visible in the Lambda reconstruction in that figure.

This also means that any null function for the interior problem must be smooth in the ROI since its Radon data are zero (e.g., smooth) for lines meeting the ROI.

Microlocal Analysis of Conical Tilt Electron Microscope Tomography (ET)

Using the notation of section "Algorithms in Conical Tilt ET," let

$$S_\alpha = \{(\sin(\alpha)\cos(\varphi), \sin(\alpha)\sin(\varphi), \cos(\alpha)) \ : \ \varphi \in [0, 2\pi]\}$$
$$C_\alpha = \mathbb{R}S_\alpha \, .$$

Using Theorem 10, one writes the operators $\mathcal{L}_\Delta$ (see (28)) and $\mathcal{L}_S$ (see (29)) using the convolution

$$f * I(x) = \int_{y \in C_\alpha} f(x+y) \frac{1}{\|y\|} \mathrm{d}y$$

where $\mathrm{d}y$ is the surface area measure on the cone C_α. In particular the conclusion of that theorem is

$$\begin{aligned} \mathcal{P}_S^* \mathcal{P}_S f &= f * I \\ \mathcal{L}_\Delta f &= (-\Delta)(f * I) \\ \mathcal{L}_S f &= \left(-\Delta + \csc^2(\alpha) \frac{\partial^2}{\partial z^3}\right) f * I \, . \end{aligned}$$

As one can see from the reconstruction in the $x-y$ plane, Fig. 4, there are circular artifacts. In the reconstruction in the vertical plane $x = -y$, Fig. 5, there are streak artifacts coming off from each of the balls at a 45° angle.

These artifacts can be understood intuitively as follows. Let f be the characteristic function of a ball B. Then, $f * I(x)$ integrates f over the cone $x + C_{\pi/4}$. When this cone is tangent to B at a point besides x, it is normal to a singularity of f at this point. This singularity will cause a singularity of $f * I$ at x. The reason is, when x is perturbed, the cone moves in and out of B, so the integral changes from 0 (when the cone is disjoint from B) to nonzero values as the cone intersects B. A calculation will convince one that the singularity is a discontinuity in the first derivative. Because $\mathcal{L}_\Delta$ is a derivative of $f * I$, that singularity is accentuated in $\mathcal{L}_\Delta f$ at x. One can see this phenomenon from the artifacts in $\mathcal{L}_\Delta$ reconstructions in section "Electron Microscope Tomography (ET) Over Arbitrary Curves." In the $x-y$ plane, the artifacts are circular shadows from the disks outside of the plane; in the $y = -x$ plane, the artifacts follow along the generating lines of the cone. However, this is not a rigorous explanation since it does not apply to arbitrary functions. It also does not explain why the $\mathcal{L}_\Delta$ reconstruction has apparent singularities, but the $\mathcal{L}_S$ reconstruction seems not to.

The next theorem explains this using microlocal analysis. To state the theorem, recall the definition of conormal bundle of a submanifold $B \subset \mathbb{R}^3$ as

$$N^*(B) = \{(y, \xi \mathrm{d}x) \,:\, y \in B, \xi \text{ is normal to } B \text{ at } y\} \,.$$

Theorem 18 ([23]). *The conical tilt ET operator $\mathcal{P}_S$ is an elliptic Fourier integral operator. Let $f \in \mathcal{E}'(\mathbb{R}^3)$ and let $\alpha \in (0, \pi/2)$. Let $\mathcal{L}$ be either $\mathcal{L}_\Delta$ or $\mathcal{L}_S$. Then,*

$$\mathrm{WF}(\mathcal{L}(f)) \subset (\mathrm{WF}(f) \cap \mathcal{V}) \cup \mathcal{A}(f)$$

where

$$\mathcal{V} = \mathbb{R}^3 \times \left\{\eta \in \mathbb{R}^3 \setminus \mathbf{0} \,:\, \exists \omega \in S_\alpha,\ \eta \cdot \omega = 0\right\}$$

represents the set of possible visible singularities. The added artifacts come from

$$\mathcal{A}(f) = \{(x, \xi \mathrm{d}x) \,:\, \exists\, (y, \xi \mathrm{d}x) \in N^*(x + C_\alpha) \cap \mathrm{WF}(f)\} \,.$$

Furthermore, the added artifacts in $\mathcal{A}(f)$ are more pronounced (stronger in Sobolev scale) in $\mathcal{L}_\Delta f$ than in $\mathcal{L}_S f$.

Here $\mathcal{V}$ is the set of visible singularities, those from f that should appear in $\mathcal{L}f$. The set $\mathcal{A}(f)$ consists of added artifacts, those caused by wavefront of f normal to $x + C_\alpha$ at points besides x. In Fig. 4, the added artifacts in the $\mathcal{L}_\Delta$ reconstruction are exactly those that come from cones $x + C_\alpha$ that are tangent to boundaries of disks in $\mathrm{supp}(f)$, and this gives the same conclusions as the heuristic description given above Theorem 18.

The final statement in the theorem follows from the fact that $\mathcal{L}_\Delta$ is in a class of singular FIOs (so-called $I^{p,l}$ classes studied in [31,32,39,64]) and $\mathcal{L}_S$ is in a better behaved class. These classes of singular FIOs also come up in radar imaging, as discussed in the next section.

The set of visible singularities $\mathcal{V}$ is larger for conical tilt ET than for the standard data acquisition geometry, single axis tilt ET, and this is one reason to take conical tilt data, even though acquiring data can be technically more difficult.

The proof of this theorem uses the Hörmander-Sato Lemma (Theorem 16). One first calculates the canonical relation, C, and then one calculates $C^t \circ C$ and shows $C^t \circ C$ has two parts. One part generates $\mathcal{V}$, the set of visible singularities, singularities of f that will be visible in $\mathcal{L}_\Delta f$. The second part generates the set of added artifacts, $\mathcal{A}(f)$.

Deep results in [31] are used to explain why $\mathcal{L}_S$ suppresses the added singularities better than $\mathcal{L}_\Delta$. This can be explained intuitively as follows. If an added singularity at x comes from a point $(y, \xi \mathrm{d}x) \in N^*(x + C_\alpha) \cap \mathrm{WF}(f)$, $\mathcal{D}_S$ takes a derivative corresponding to a derivative in x that is normal to ξ (so it does not increase the order of this singularity) and Δ takes a derivative *in* the direction of this singularity (so it does increase the order of this singularity).

This conical tilt transform is an *admissible Radon transform* [27]. The microlocal analysis of such transforms was first studied in [36], and the microlocal properties of these transforms were completely analyzed in a very general setting in [30] including general results for backprojection that are related to Theorem 18.

Quinto and Rullgård have proven similar results for a curvilinear Radon transform for which the more effective differential operator (analogous to $\mathcal{D}_S$) decreases the strength of artifacts only locally [81].

SAR Imaging

In this section, some recent results on SAR imaging will be described from the viewpoint of microlocal analysis. The notation is from section "Synthetic-Aperture Radar Imaging." Recall that the forward operator under consideration is

$$\mathcal{P}V(s,t) = \int e^{-i\omega(t-(\|x-\gamma_T(s)\|+\|x-\gamma_R(s)\|)/c_0)} A(s,t,x,\omega)V(x)\mathrm{d}x\mathrm{d}\omega. \tag{57}$$

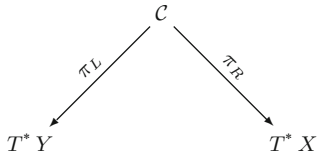
In this section $\omega \in \mathbb{R}$. The canonical relation wherever it is well defined is given as follows. This is an important set that relates the singularities of the object V to that of the data $\mathcal{P}V$:

$$\begin{aligned} C = \Bigg\{ & s,t,-\omega\left(\left(\frac{x-\gamma_T(s)}{\|x-\gamma_T(s)\|}\cdot\gamma_T'(s) + \frac{x-\gamma_R(s)}{\|x-\gamma_R(s)\|}\cdot\gamma_R'(s)\right)\mathrm{d}s + \mathrm{d}t\right); \\ & x_1, x_2, -\omega\left(\frac{x-\gamma_T(s)}{\|x-\gamma_T(s)\|} + \frac{x-\gamma_R(s)}{\|x-\gamma_R(s)\|}\right)\mathrm{d}x \\ & : c_0 t = \|x-\gamma_T(s)\| + \|x-\gamma_R(s)\|, \omega \neq 0 \Bigg\}. \end{aligned} \tag{58}$$

Note that (s,x,ω) is a global parametrization for C. Let us denote

$$Y = \{(s,t) \in (0,\infty)\times(0,\infty)\}$$

and $\{x_1, x_2\}$ space as X. One is interested in studying the imaging operator $\mathcal{P}^*\mathcal{P}$. The standard composition calculus of FIOs, the so-called transverse intersection calculus of Hörmander [46], and the clean intersection calculus of Duistermaat and Guillemin [14, 34] and Weinstein [98] do not apply in general in these situations. Therefore, one approach to understanding the imaging operator is to study the canonical left and right projections from the canonical relation $C \subset T^*Y \times T^*X$ to T^*Y and T^*X, respectively.



(59)

In order to motivate the results that follow, let us consider the following example.

Example 12. Let us consider a simple example from SAR imaging. This example will help us explain via microlocal analysis, some of the artifacts introduced by image reconstruction operator in 6.

Assume that a colocated transmitter/receiver traverses the straight trajectory $\gamma(s) = (s, 0, h)$ with h fixed. The forward operator in this case is

$$\mathcal{P}f(s,t) = \int e^{-\mathrm{i}\omega\left(t-\frac{2}{c_0}\sqrt{(x-s)^2+y^2+h^2}\right)} A(s,t,x,y,\omega) f(x,y)\mathrm{d}x\mathrm{d}y\mathrm{d}\omega.$$

Using (51) and (52), the canonical relation of this operator is easily computed to be

$$C = \left\{ s, \frac{2}{c_0}\sqrt{(x-s)^2+y^2+h^2}, -\omega\left(\frac{2}{c_0}\frac{x-s}{\sqrt{(x-s)^2+y^2+h^2}}\mathrm{d}s + \mathrm{d}t\right); \right.$$
$$\left. x, y, -\frac{2\omega}{c_0}\left(\frac{x-s}{\sqrt{(x-s)^2+y^2+h^2}}\mathrm{d}x + \frac{y}{\sqrt{(x-s)^2+y^2+h^2}}\mathrm{d}y\right)\right\}.$$

Now using Hörmander-Sato Lemma, one sees that

$$\mathrm{WF}(\mathcal{P}^*\mathcal{P}) \subset \left\{ x, y, \frac{2\omega}{c_0}\left(\frac{x-s}{\sqrt{(x-s)^2+y^2+h^2}}\mathrm{d}x + \frac{y}{\sqrt{(x-s)^2+y^2+h^2}}\mathrm{d}y\right); \right.$$
$$z, w, \frac{2\omega}{c_0}\left(\frac{z-s}{\sqrt{(z-s)^2+w^2+h^2}}\mathrm{d}z + \frac{w}{\sqrt{(z-s)^2+w^2+h^2}}\mathrm{d}w\right):$$
$$\sqrt{(x-s)^2+y^2+h^2} = \sqrt{(z-s)^2+w^2+h^2} \text{ and}$$
$$\left. \frac{x-s}{\sqrt{(x-s)^2+y^2+h^2}} = \frac{z-s}{\sqrt{(z-s)^2+w^2+h^2}}, \omega \neq 0 \right\}.$$

This implies that $(z, w) = (x, y)$ or $(z, w) = (x, -y)$. The first equality contributes to the diagonal relation of the wavefront set of $\mathcal{P}^*\mathcal{P}$, while the second contributes to the relation formed by reflection about the x-axis. In other words,

$$\mathrm{WF}(\mathcal{P}^*\mathcal{P}) \subset \triangle \cup G, \text{ where } \triangle = \{(x, y, \xi\mathrm{d}x + \eta\mathrm{d}y : x, y, \xi\mathrm{d}x + \eta\mathrm{d}y)\}$$

and $G = \{(x, y, \xi \mathrm{d}x + \eta \mathrm{d}y; x, -y, \xi \mathrm{d}x - \eta \mathrm{d}y)\}$.

The presence of the set G as in the above example (the non-diagonal part) indicates that the imaging operator introduces artifacts in the reconstructed image. A detailed study of the class of distributions as in the example above, called $I^{p,l}$ classes, was introduced in [31, 32, 39, 64]. These classes of distributions have come up in the study of several imaging problems including X-ray, seismic, SAR, and electron tomography [1, 13, 21–24, 30–32, 54, 69]. In instances where the imaging operator introduces artifacts, it is of interest to determine whether the artifacts are of the same strength in a suitable sense as that of the true singularities and whether the artifacts can be suppressed or displaced from the true singularities. These questions are answered in the references given above.

Monostatic SAR Imaging

In monostatic SAR imaging, the transmitter and receiver are located at the same point. In other words, $\gamma_T(s) = \gamma_R(s)$. Nolan and Cheney in [70] investigated the microlocal properties of the forward operator $\mathcal{P}$ and the associated image reconstruction operator $\mathcal{P}^*\mathcal{P}$. Using microlocal tools, synthetic-aperture inversion in the presence of noise and clutter was done in [102]. Other imaging methods in the context of SAR imaging, again using microlocal tools, were considered in [99, 101]. The forward operator $\mathcal{P}$ was further investigated by Felea in [19], and she made a detailed analysis of the image reconstruction operator for various flight trajectories. Felea in [19] showed that for $\gamma(s) = (s, 0, h)$ with $h > 0$ fixed, the operator $\mathcal{P}^*\mathcal{P}$ belongs to $I^{2m,0}(\Delta, G)$ where $\Delta = \{(x, \xi, x, \xi) \in T^*X \times T^*X \setminus \mathbf{0}\}$ and G is the graph of the function $\chi(x_1, x_2, \xi_1, \xi_2) = (x_1, -x_2, \xi_1, -\xi_2)$. If $\gamma(s) = (\cos s, \sin s, h)$, it was shown in [20] that $\mathcal{P}^*\mathcal{P} \in I^{2m,0}(\Delta, G)$, where G is a 2-sided fold. Mappings with singularities (such as folds and blowdowns) are defined in [29, 37]. Furthermore, in [20], Felea showed that in some instances such as the flight trajectory being circular, the artifact singularities of the same strength as the true singularities can be displaced far away from the true singularities, and those that are not displaced are of lesser strength compared to the true singularities. In [91], the authors show that cancelation of singularities, that is, only certain singularities are recoverable, can occur even in curved flight paths.

Bistatic SAR Imaging

Some recent results by the authors and their collaborators investigating the microlocal properties of transforms that appear in bistatic SAR imaging are now described. For related work, the reader is referred to [100].

Common Offset Bistatic SAR Imaging

In common offset SAR imaging, the transmitter and receiver travel in a straight line offset by a constant distance at all times. More precisely, let

$$\gamma_T(s) = (s + \alpha, 0, h) \text{ and } \gamma_R(s) = (s - \alpha, 0, h)$$

be the trajectories of the transmitter and receiver, respectively, with α and h fixed positive quantities. A detailed microlocal analysis of associated forward operator $\mathcal{P}$ and the imaging operator $\mathcal{P}^*\mathcal{P}$ was done in [54]. The authors obtained the following results analogous to the ones obtained by Nolan and Cheney in [70] and Felea in [19] for the monostatic case.

Theorem 19 ([54]). *Let $\gamma_T(s) = (s + \alpha, 0, h)$ and $\gamma_R(s) = (s - \alpha, 0, h)$ where $\alpha > 0, h > 0$ are fixed. The operator $\mathcal{P}$ defined in (57) is an FIO. The canonical relation C associated to $\mathcal{P}$ defined in (58) satisfies the following: the projections π_L and π_R defined in (59) are a fold and blowdown, respectively.*

Theorem 20 ([54]). *Let $\mathcal{P}$ be defined with γ_T and γ_R given in Theorem 19. Then $\mathcal{P}^*\mathcal{P} \in I^{3,0}(\Delta, G)$, where Δ is the diagonal relation and G is the graph of the map $\chi(x_1, x_2, \xi_1, \xi_2) = (x_1, -x_2, \xi_1, -\xi_2)$.*

One important consequence of this result is that the artifacts given by the graph G of the map χ have the same strength in a Sobolev sense as that of the true singularities given by the diagonal relation Δ.

Common Midpoint SAR Imaging

In common midpoint SAR imaging, the transmitter and receiver travel in a straight line at a constant height above the ground at equal speeds away from a common midpoint. The trajectories of the transmitter and receiver for the common midpoint geometry considered here are

$$\gamma_T(s) = (s, 0, h) \text{ and } \gamma_R(s) = (-s, 0, h). \tag{60}$$

A detailed microlocal analysis of the forward operator (58) associated to γ_T and γ_R and the imaging operator $\mathcal{P}^*\mathcal{P}$ was done in [1]. In contrast to the results in [19, 54, 70], here the canonical relation C associated to $\mathcal{P}$ is a 4-1 relation, and this is reflected in the fact the canonical left and right projections π_L and π_R drop rank on a union of two disjoint sets. More precisely, one obtains the following results for the forward operator and the imaging operator, respectively.

Theorem 21 ([1]). *Let $\mathcal{P}$ be as in (57) with the trajectories given by (60). Then $\mathcal{P}$ is an FIO and the canonical relation associated to $\mathcal{P}$ defined in (58) has global parametrization*

$$(0, \infty) \times \left(\mathbb{R}^2 \setminus 0\right) \times (\mathbb{R} \setminus 0) \ni (s, x_1, x_2, \omega) \mapsto C,$$

and it satisfies the following: the left and right projections π_L and π_R drop rank simply by one on a set $\Sigma = \Sigma_1 \cup \Sigma_2$ where in the coordinates (s, x, ω), $\Sigma_1 = \{(s, x_1, 0, \omega) : s > 0, |x_1| > \epsilon', \omega \neq 0\}$ and $\Sigma_2 = \{(s, 0, x_2, \omega) : s > 0, |x_2| > \epsilon', \omega \neq 0\}$ for $0 < \epsilon'$ small enough. The canonical relation C associated to $\mathcal{P}$

satisfies the following: the projections π_L and π_R defined in (59) are a fold and blowdown, respectively, along Σ.

Theorem 22 ([1]). *Let $\mathcal{P}$ be as in (57) with the trajectories given by (60). Then $\mathcal{P}^*\mathcal{P}$ can be decomposed into a sum belonging to $I^{2m,0}(\Delta, G_1) + I^{2m,0}(\Delta, G_2) + I^{2m,0}(G_1, G_3) + I^{2m,0}(G_2, G_3)$, where G_i for $i = 1, 2, 3$ are the graphs of the following functions χ_i for $i = 1, 2, 3$ on T^*X:*

$$\chi_1(x,\xi) = (x_1, -x_2, \xi_1, -\xi_2), \quad \chi_2(x,\xi) = (-x_1, x_2, -\xi_1, \xi_2) \text{ and } \chi_3 = \chi_1 \circ \chi_2.$$

As with the common offset case, the artifacts given by graphs of the maps χ_1, χ_2, and χ_3 have the same strength in a Sobolev sense as that of the true singularities given by the diagonal relation Δ.

6 Conclusion

Finally, some important themes of this chapter will be highlighted.

Microlocal analysis can help understand strengths and limitations of tomographic reconstruction methods. For limited data X-ray tomography, microlocal analysis was used in this chapter to show which singularities of functions will be visible depending on the data, and reconstructions in this chapter illustrate the paradigm.

Each of the reconstruction methods described in this chapter is of the form

$$\mathcal{L} = \mathcal{P}^*\mathcal{D}\mathcal{P}$$

where the forward operator (operator modeling the tomography problem) $\mathcal{P}$ is a Fourier integral operator and $\mathcal{P}^*$ is an adjoint and $\mathcal{D}$ is a pseudodifferential operator. In SAR imaging, $\mathcal{D} = \mathrm{Id}$ is the identity operator and the reconstruction method is $\mathcal{P}^*\mathcal{P}$ – the normal operator. Since the operator $\mathcal{D}$ is a differential or pseudodifferential operator, it does not add to the wavefront set of $\mathcal{P}f$, $\mathrm{WF}(\mathcal{D}\mathcal{P}f) \subset \mathrm{WF}(\mathcal{P}f)$. If $\mathcal{D}$ is elliptic (on the range of $\mathcal{P}$, which is true in the cases considered here), $\mathrm{WF}(\mathcal{D}\mathcal{P}f) = \mathrm{WF}(\mathcal{P}f)$. Then, one needs to understand what $\mathcal{P}^*$ does, and this is determined by the structure of $C^t \circ C$ by the Hörmander Sato Lemma (Theorem 16):

$$\mathrm{WF}\left(\mathcal{P}^*\left(\mathcal{D}\mathcal{P}f\right)\right) \subset \left(C^t \circ C\right) \circ \mathrm{WF}(f)\,.$$

However, in limited angle tomography, because $\mathcal{R}^*_{L,\mathrm{lim}}$ is not a standard FIO, $\mathcal{R}^*_{L,\mathrm{lim}}\mathcal{D}\mathcal{R}_L$ adds singularities along lines at the ends of the limited angular range.

In the case of SAR and conical tilt ET, $C^t \circ C$ is more complicated; it includes Δ and another set. In conical tilt ET, this second set generates $\mathcal{A}(f)$, the extra artifacts given in Theorem 18. Also, for conical tilt ET, that theorem implies that a well-chosen differential operator ($(-\mathcal{D}_S)$ rather than $(-\Delta_S)$) decreases the strength of the added singularities; they are visible if one looks carefully at the reconstructions, but they are smoother than when $(-\Delta_S)$ is used.

The only exact reconstruction method presented in this chapter is FBP (Theorem 9). The other algorithms, such as Lambda CT, involve differential operators and backprojection. Microlocal analysis was used to demonstrate that they do recover many (or all) singularities of the object.

Recovering singularities does not recover the object. So, these algorithms are not useful when one needs density values, such as in distinguishing tumor cells from benign cells in diagnostic radiology. However, in many cases (e.g., ET or industrial nondestructive evaluation) one is interested in the shapes of regions, not actual density values, so knowing the location of singularities is useful. The algorithm must be designed so that it clearly shows singularities in the reconstruction. For example, an algorithm that turns jump discontinuities in the object into discontinuities of a derivative in the reconstruction might not provide a clear picture of the object. Lambda CT and the algorithms given here for conical tilt ET actually accentuate singularities; they make the singularities more apparent since they are operators of order one (like a first derivative). The implementation smooths out the derivative since it uses numerical differentiation.

In conical tilt ET and SAR the reconstruction methods, that included a backprojection (adjoint, $\mathcal{P}^*$), produce added artifacts (as shown in Figs. 5 and 6), and these were explained using microlocal analysis.

Microlocal analysis will not make a bad algorithm good, but it can show that some limitations in reconstruction quality are intrinsic to the underlying tomographic problem. It can point to where reconstruction methods need to be regularized more strongly because of intrinsic instability in the specific tomography problem. In summary, microlocal analysis can be used to understand practical and theoretical issues in tomography.

Acknowledgments Both authors thank the American Institute of Mathematics and their colleagues Gaik Ambartsoumian, Raluca Felea, and Clifford Nolan in an American Institute of Mathematics (AIM) SQuaREs program for discussions at AIM on microlocal analysis and radar imaging that informed this work. They appreciate Jürgen Frikel's careful reading of the chapter. The authors thank the Mittag Leffler Institute for the congenial atmosphere as they worked on some of the research presented in this chapter. The second named author thanks Jan Boman, Alfred Louis, Frank Natterer, and many other friends and colleagues for important discussions about tomography and microlocal analysis over the years. Both authors thank Birsen Yazıcı for interesting discussions.

The first named author was partially supported by NSF grant DMS 1109417. Additionally, he benefited from the support of Airbus Group Corporate Foundation Chair in "Mathematics of Complex Systems" established at TIFR CAM and ICTS TIFR, Bangalore, India, and from a German DAAD Research Stays grant. The second named author was partially supported by NSF grant DMS 1311558.

Cross-References

▸ Compressive Sensing
▸ Electrical Impedance Tomography
▸ Imaging in Random Media

- Inverse Scattering
- Large-Scale Inverse Problems in Imaging
- Level Set Methods for Structural Inversion and Image Reconstruction
- Linear Inverse Problems
- Mathematics of Electron Tomography
- Mathematical Methods of Optical Coherence Tomography
- Mathematical Methods in PET and Spect Imaging
- Mathematics of Photoacoustic and Thermoacoustic Tomography
- Mumford and Shah Model and Its Applications to Image Segmentation and Image Restoration
- Optical Imaging
- Photoacoustic and Thermoacoustic Tomography: Image Formation Principles
- Regularization Methods for Ill-Posed Problems
- Sampling Methods
- Synthetic Aperture Radar Imaging
- Tomography
- Total Variation in Imaging
- Wave Phenomena

References

1. Ambartsoumian, G., Felea, R., Krishnan, V.P., Nolan, C., Quinto, E.T.: A class of singular Fourier integral operators in synthetic aperture radar imaging. J. Funct. Anal. **264**(1), 246–269 (2013). doi:10.1016/j.jfa.2012.10.008, http://dx.doi.org/10.1016/j.jfa.2012.10.008
2. Anastasio, M.A., Zou, Y., Sidky, E.Y., Pan, X.: Local cone-beam tomography image reconstruction on chords. J. Opt. Soc. Am. A **24**, 1569–1579 (2007)
3. Bates, R., Lewitt, R.: Image reconstruction from projections: I: general theoretical considerations, II: projection completion methods (theory), III: Projection completion methods (computational examples). Optik **50**, I: 19–33, II: 189–204, III: 269–278 (1978)
4. Bidwell, S.: Limited angle tomography and microlocal analysis. Tech. rep., Tufts University, senior Honors Thesis with Highest Thesis Honors (2012)
5. Boman, J.: Helgason's Support Theorem for Radon Transforms – A New Proof and a Generalization. Lecture Notes in Mathematics, vol. 1497, pp. 1–5. Springer, Berlin/New York (1991)
6. Boman, J., Quinto, E.T.: Support theorems for real analytic Radon transforms. Duke Math. J. **55**, 943–948 (1987)
7. Cheney, M., Borden, B.: Fundamentals of Radar Imaging, CBMS-NSF Regional Conference Series in Applied Mathematics, vol. 79. Society for Industrial and Applied Mathematics (SIAM), Philadelphia (2009). doi:10.1137/1.9780898719291, http://dx.doi.org/10.1137/1.9780898719291
8. Cheney, M., Borden, B.: Problems in synthetic-aperture radar imaging. Inverse Probl. **25**(12), 123005, 18 (2009). doi:10.1088/0266-5611/25/12/123005, http://dx.doi.org/10.1088/0266-5611/25/12/123005
9. Cheney, M., Bordon, B.: Synthetic aperture radar imaging. In: Scherzer, O. (ed.) The Handbook of Mathematical Methods in Imaging, 2nd edn. Springer, New York (2014)
10. Cormack, A.M.: Representation of a function by its line integrals with some radiological applications. J. Appl. Phys. **34**, 2722–2727 (1963)

11. Cormack, A.M.: Representation of a function by its line integrals with some radiological applications II. J. Appl. Phys. **35**, 2908–2913 (1964)
12. Davison, M., Grünbaum, F.: Tomographic reconstruction with arbitrary directions. Commun. Pure Appl. Math. **34**, 77–120 (1981)
13. de Hoop, M.V.: Microlocal analysis of seismic inverse scattering. In: Inside Out: Inverse Problems and Applications. Mathematical Sciences Research Institute Publications, vol. 47, pp. 219–296. Cambridge University Press, Cambridge (2003)
14. Duistermaat, J.J., Guillemin, V.W.: The spectrum of positive elliptic operators and periodic bicharacteristics. Invent. Math. **29**(1), 39–79 (1975)
15. Faridani, A.: Tomography and sampling theory. In: Olafsson, G., Quinto, E.T. (eds.) The Radon Transform and Applications to Inverse Problems. American Mathematical Society, Providence. AMS Proceedings of Symposia in Applied Mathematics (2006)
16. Faridani, A., Ritman, E.L.: High-resolution computed tomography from efficient sampling. Inverse Probl. **16**, 635–650 (2000)
17. Faridani, A., Ritman, E.L., Smith, K.T.: Local tomography. SIAM J. Appl. Math. **52**(2), 459–484 (1992). doi:10.1137/0152026, http://dx.doi.org/10.1137/0152026
18. Faridani, A., Finch, D., Ritman, E., Smith, K.T.: Local tomography, II. SIAM J. Appl. Math. **57**, 1095–1127 (1997)
19. Felea, R.: Composition of Fourier integral operators with fold and blowdown singularities. Commun. Partial Differ. Equ. **30**(10–12), 1717–1740 (2005). doi:10.1080/03605300500299968, http://dx.doi.org/10.1080/03605300500299968
20. Felea, R.: Displacement of artefacts in inverse scattering. Inverse Probl. **23**(4), 1519–1531 (2007). doi:10.1088/0266-5611/23/4/009, http://dx.doi.org/10.1088/0266-5611/23/4/009
21. Felea, R., Greenleaf, A.: An FIO calculus for marine seismic imaging: folds and cross caps. Commun. Partial Differ. Equ. **33**(1–3), 45–77 (2008). doi:10.1080/03605300701318716, http://dx.doi.org/10.1080/03605300701318716
22. Felea, R., Greenleaf, A.: Fourier integral operators with open umbrellas and seismic inversion for cusp caustics. Math. Res. Lett. **17**(5), 867–886 (2010). doi:10.4310/MRL.2010.v17.n5.a6, http://dx.doi.org/10.4310/MRL.2010.v17.n5.a6
23. Felea, R., Quinto, E.T.: The microlocal properties of the local 3-D spect operator. SIAM J. Math. Anal. **43**(3), 659–674 (2011)
24. Felea, R., Greenleaf, A., Pramanik, M.: An FIO calculus for marine seismic imaging, II: Sobolev estimates. Math. Ann. **352**(2), 293–337 (2012). doi:10.1007/s00208-011-0644-5, http://dx.doi.org/10.1007/s00208-011-0644-5
25. Finch, D.V., Lan, I.R., Uhlmann, G.: Microlocal analysis of the restricted X-ray transform with sources on a curve. In: Uhlmann, G. (ed.) Inside Out, Inverse Problems and Applications. MSRI Publications, vol. 47, pp. 193–218. Cambridge University Press, Cambridge/New York (2003)
26. Frikel, J., Quinto, E.T.: Characterization and reduction of artifacts in limited angle tomography. Inverse Probl. **29**, 125007 (2013)
27. Gelfand, I.M., Graev, M.I.: Integral transformations connected with straight line complexes in a complex affine space. Sov. Math. Dokl. **2**, 809–812 (1961)
28. Gelfand, I.M., Graev, M.I., Vilenkin, N.Y.: Generalized Functions, vol. 5. Academic, New York (1966)
29. Golubitsky, M., Guillemin, V.: Stable Mappings and Their Singularities. Graduate Texts in Mathematics, vol. 14. Springer, New York (1973)
30. Greenleaf, A., Uhlmann, G.: Non-local inversion formulas for the X-ray transform. Duke Math. J. **58**, 205–240 (1989)
31. Greenleaf, A., Uhlmann, G.: Composition of some singular Fourier integral operators and estimates for restricted X-ray transforms. Ann. Inst. Fourier (Grenoble) **40**(2), 443–466 (1990). http://www.numdam.org/item?id=AIF_1990__40_2_443_0
32. Greenleaf, A., Uhlmann, G.: Estimates for singular Radon transforms and pseudodifferential operators with singular symbols. J. Funct. Anal. **89**(1), 202–232 (1990). doi:10.1016/0022-1236(90)90011-9, http://dx.doi.org/10.1016/0022-1236(90)90011-9

33. Grigis, A., Sjöstrand, J.: Microlocal Analysis for Differential Operators: An Introduction. London Mathematical Society Lecture Note Series, vol. 196. Cambridge University Press, Cambridge (1994)
34. Guillemin, V.: Clean intersection theory and Fourier integrals. In: Fourier Integral Operators and Partial Differential Equations. Colloque International, Université de Nice, Nice, 1974. Lecture Notes in Mathematics, vol. 459, pp. 23–35. Springer, Berlin (1975)
35. Guillemin, V.: Some remarks on integral geometry. Technical report, MIT (1975)
36. Guillemin, V.: On some results of Gelfand in integral geometry. Proc. Symp. Pure Math. **43**, 149–155 (1985)
37. Guillemin, V.: Cosmology in $(2+1)$-Dimensions, Cyclic Models, and Deformations of $M_{2,1}$. Annals of Mathematics Studies, vol. 121. Princeton University Press, Princeton (1989)
38. Guillemin, V., Sternberg, S.: Geometric Asymptotics. Mathematical Surveys, vol. 14. American Mathematical Society, Providence (1977)
39. Guillemin, V., Uhlmann, G.: Oscillatory integrals with singular symbols. Duke Math. J. **48**(1), 251–267 (1981). http://projecteuclid.org/getRecord?id=euclid.dmj/1077314493
40. Hahn, M.G., Quinto, E.T.: Distances between measures from 1-dimensional projections as implied by continuity of the inverse Radon transform. Zeitschrift Wahrscheinlichkeit **70**, 361–380 (1985)
41. Helgason, S.: The Radon transform on Euclidean spaces, compact two-point homogeneous spaces and Grassman manifolds. Acta Math. **113**, 153–180 (1965)
42. Helgason, S.: Support of Radon transforms. Adv. Math. **38**, 91–100 (1980)
43. Helgason, S.: Integral Geometry and Radon Transforms. Springer, New York (2011). doi:10.1007/978-1-4419-6055-9
44. Herman, G.: Tomography. In: Scherzer, O. (ed.) The Handbook of Mathematical Methods in Imaging, 2nd edn. Springer, New York (2014)
45. Hertle, A.: Continuity of the Radon transform and its inverse on Euclidean space. Mathematische Zeitschrift **184**, 165–192 (1983)
46. Hörmander, L.: Fourier integral operators, I. Acta Math. **127**, 79–183 (1971)
47. Hörmander, L.: The Analysis of Linear Partial Differential Operators. I. Classics in Mathematics. Springer, Berlin (2003). Distribution theory and Fourier analysis, Reprint of the second (1990) edition [Springer, Berlin; MR1065993 (91m:35001a)]
48. Horne, A., Yates, G.: Bistatic synthetic aperture radar. In: 2002 International Radar Conference, pp. 6–10(4) (2002). doi:10.1049/cp:20020238, http://digital-library.theiet.org/content/conferences/10.1049/cp_20020238
49. Katsevich, A.I.: Local tomography for the limited-angle problem. J. Math. Anal. Appl. **213**, 160–182 (1997)
50. Katsevich, A.I.: An improved exact filtered backprojection algorithm for spiral computed tomography. Adv. Appl. Math. **32**, 681–697 (2004)
51. Katsevich, A.: Improved cone beam local tomography. Inverse Probl. **22**, 627–643 (2006)
52. Katsevich, A., Ramm, A.: The Radon Transform and Local Tomography. CRC Press, Boca Raton (1996)
53. Krishnan, V.P.: A support theorem for the geodesic ray transform on functions. J. Fourier Anal. Appl. **15**(4), 515–520 (2009). doi:10.1007/s00041-009-9061-5
54. Krishnan, V.P., Quinto, E.T.: Microlocal aspects of common offset synthetic aperture radar imaging. Inverse Probl. Imaging **5**(3), 659–674 (2011). doi:10.3934/ipi.2011.5.659, http://dx.doi.org/10.3934/ipi.2011.5.659
55. Krishnan, V.P., Levinson, H., Quinto, E.T.: Microlocal analysis of elliptical Radon transforms with foci on a line. In: Sabadini, I., Struppa, D.C. (eds.) The Mathematical Legacy of Leon Ehrenpreis. Springer Proceedings in Mathematics, vol. 16, pp. 163–182. Springer, Berlin/New York (2012)
56. Kuchment, P., Lancaster, K., Mogilevskaya, L.: On local tomography. Inverse Probl. **11**, 571–589 (1995)
57. Kurusa, Á.: Support theorems for totally geodesic Radon transforms on constant curvature spaces. Proc. Am. Math. Soc. **122**(2), 429–435 (1994)

58. Lissianoi, S.: On stability estimates in the exterior problem for the Radon transform. In: Quinto, E., Ehrenpreis, L., Faridani, A., Gonzalez, F., Grinberg, E. (eds.) Tomography, Impedance Imaging, and Integral Geometry, South Hadley, 1993. Lectures in Applied Mathematics, vol. 30, pp. 143–147. American Mathematical Society, Providence (1994)
59. Louis, A.K.: Analytische Methoden in der Computer Tomographie. Universität Münster, habilitationsschrift (1981)
60. Louis, A.K.: Ghosts in tomography, the null space of the Radon transform. Math. Methods Appl. Sci. **3**, 1–10 (1981)
61. Louis, A.K.: Incomplete data problems in X-ray computerized tomography I. Singular value decomposition of the limited angle transform. Numerische Mathematik **48**, 251–262 (1986)
62. Louis, A.K., Maaß, P.: Contour reconstruction in 3-D X-Ray CT. IEEE Trans. Med. Imaging **12**(4), 764–769 (1993)
63. Louis, A.K., Rieder, A.: Incomplete data problems in X-ray computerized tomography II. Truncated projections and region-of-interest tomography. Numerische Mathematik **56**, 371–383 (1986)
64. Melrose, R.B., Uhlmann, G.A.: Lagrangian intersection and the Cauchy problem. Commun. Pure Appl. Math. **32**(4), 483–519 (1979). doi:10.1002/cpa.3160320403, http://dx.doi.org/10.1002/cpa.3160320403
65. Natterer, F.: Efficient implementation of 'optimal' algorithms in computerized tomography. Math. Methods Appl. Sci. **2**, 545–555 (1980)
66. Natterer, F.: The Mathematics of Computerized Tomography. Classics in Applied Mathematics, vol. 32. Society for Industrial and Applied Mathematics (SIAM), Philadelphia (2001). Reprint of the 1986 original
67. Natterer, F., Wübbeling, F.: Mathematical Methods in Image Reconstruction. Monographs on Mathematical Modeling and Computation. Society for Industrial and Applied Mathematics, New York (2001)
68. Niinimaki, K., Siltanen, S., Kolehmainen, V.: Bayesian multiresolution method for local tomography in dental X-ray imaging. Phys. Med. Biol. **52**, 6663–6678 (2007)
69. Nolan, C.J.: Scattering in the presence of fold caustics. SIAM J. Appl. Math. **61**(2), 659–672 (2000). doi:10.1137/S0036139999356107, http://dx.doi.org/10.1137/S0036139999356107
70. Nolan, C.J., Cheney, M.: Microlocal analysis of synthetic aperture radar imaging. J. Fourier Anal. Appl. **10**(2), 133–148 (2004). doi:10.1007/s00041-004-8008-0, http://dx.doi.org/10.1007/s00041-004-8008-0
71. Öktem, O.: The mathematics of electron microscopy. In: Scherzer, O. (ed.) The Handbook of Mathematical Methods in Imaging, 2nd edn. Springer, New York (2014)
72. Park, J.M., Franken, E.A., Jr., Garg, M., Fajardo, L.L., Niklason, L.T.: Breast tomosynthesis: present considerations and future applications. RadioGraphics **27**, S231–S240 (2007)
73. Perry, R.M.: On reconstructing a function on the exterior of a disc from its Radon transform. J. Math. Anal. Appl. **59**, 324–341 (1977)
74. Petersen, B.: Introduction to the Fourier Transform and Pseudo-differential Operators. Pittman, Boston (1983)
75. Quinto, E.T.: The dependence of the generalized Radon transform on defining measures. Trans. Am. Math. Soc. **257**, 331–346 (1980)
76. Quinto, E.T.: Singular value decompositions and inversion methods for the exterior Radon transform and a spherical transform. J. Math. Anal. Appl. **95**, 437–448 (1983)
77. Quinto, E.T.: Tomographic reconstructions from incomplete data–numerical inversion of the exterior Radon transform. Inverse Probl. **4**, 867–876 (1988)
78. Quinto, E.T.: Support theorems for the spherical Radon transform on manifolds. Int. Math. Res. Not. **2006**, 1–17 (2006). Article ID = 67205
79. Quinto, E.T.: Local algorithms in exterior tomography. J. Comput. Appl. Math. **199**, 141–148 (2007)
80. Quinto, E.T., Öktem, O.: Local tomography in electron microscopy. SIAM J. Appl. Math. **68**, 1282–1303 (2008)

81. Quinto, E.T., Rullgård, H.: Local singularity reconstruction from integrals over curves in R^3. Inverse Probl. Imaging **7**(2), 585–609 (2013)
82. Quinto, E.T., Bakhos, T., Chung, S.: A local algorithm for Slant Hole SPECT. In: Mathematical Methods in Biomedical Imaging and Intensity-Modulated Radiation Therapy (IMRT), Centro De Georgi, Pisa. cRM Series, vol. 7, pp. 321–348 (2008)
83. Quinto, E.T., Skoglund, U., Öktem, O.: Electron lambda-tomography. Proc. Natl. Acad. Sci. USA **106**(51), 21842–21847 (2009)
84. Radon, J.: Über die Bestimmung von Funktionen durch ihre Integralwerte längs gewisser Mannigfaltigkeiten. Ber. Verh. Sach. Akad. **69**, 262–277 (1917)
85. Ramachandran, G., Lakshminarayanan, A.: Three dimensional reconstruction from radiographs and electron micrographs: applications of convolutions instead of Fourier transforms. Proc. Natl. Acad. Sci. USA **68**, 262–277 (1971)
86. Rudin, W.: Functional Analysis. McGraw-Hill Series in Higher Mathematics. McGraw-Hill Book Co., New York (1973)
87. Shepp, L.A., Kruskal, J.B.: Computerized tomography: the new medical X-ray technology. Am. Math. Mon. **85**, 420–439 (1978)
88. Shepp, L.A., Srivastava, S.: Computed tomography of PKM and AKM exit cones. AT & T Tech. J. **65**, 78–88 (1986)
89. Shubin, M.A.: Pseudodifferential Operators and Spectral Theory, 2nd edn. Springer, Berlin (2001) (Translated from the 1978 Russian original by Stig I. Andersson)
90. Smith, K.T., Solmon, D.C., Wagner, S.L.: Practical and mathematical aspects of the problem of reconstructing objects from radiographs. Bull. Am. Math. Soc. **83**, 1227–1270 (1977)
91. Stefanov, P., Uhlmann, G.: Is a curved flight path in SAR better than a straight one? SIAM J. Appl. Math. **73**(4), 1596–1612 (2013)
92. Strichartz, R.S.: Radon inversion–variations on a theme. Am. Math. Mon. **89**, 377–384 (1982)
93. Taylor, M.E.: Pseudodifferential Operators. Princeton Mathematical Series, vol. 34. Princeton University Press, Princeton (1981)
94. Trèves, F.: Introduction to Pseudodifferential and Fourier Integral Operators. The University Series in Mathematics, vol. 1. Pseudodifferential operators. Plenum Press, New York (1980).
95. Trèves, F.: Introduction to Pseudodifferential and Fourier Integral Operators. The University Series in Mathematics, vol. 2. Fourier integral operators. Plenum Press, New York (1980).
96. Vainberg, E., Kazak, I.A., Kurozaev, V.P.: Reconstruction of the internal three-dimensional structure of objects based on real-time integral projections. Sov. J. Nondestruct. Test. **17**, 415–423 (1981)
97. Vaughan, C.L.: Imagining The Elephant: A Biography of Allan Macleod Cormack. Imperial College Press, London (2008)
98. Weinstein, A.: On Maslov's quantization condition. In: Fourier Integral Operators and Partial Differential Equations. Colloque International, Université de Nice, Nice, 1974. Lecture Notes in Mathematics, vol. 459, pp. 341–372. Springer, Berlin (1975)
99. Yarman, C.E., Yazıcı, B.: Synthetic aperture hitchhiker imaging. IEEE Trans. Image Process. **17**(11), 2156–2173 (2008). doi:10.1109/TIP.2008.2002839, http://dx.doi.org/10.1109/TIP.2008.2002839
100. Yarman, C.E., Yazıcı, B., Cheney, M.: Bistatic synthetic aperture radar imaging for arbitrary flight trajectories. IEEE Trans. Image Process. **17**(1), 84–93 (2008)
101. Yarman, C.E., Wang, L., Yazıcı, B.: Doppler synthetic aperture hitchhiker imaging. Inverse Probl. **26**(6), 065006, 26 (2010). doi:10.1088/0266-5611/26/6/065006, http://dx.doi.org/10.1088/0266-5611/26/6/065006
102. Yazici, B., Cheney, M., Yarman, C.: Synthetic-aperture inversion in the presence of noise and clutter. Inverse Probl. **22**, 1705–1729 (2006)
103. Ye, Y., Yu, H., Wang, G.: Cone beam pseudo-lambda tomography. Inverse Probl. **23**, 203–215 (2007)
104. Ye, Y., Yu, H., Wang, G.: Exact Interior Reconstruction from Truncated Limited-Angle Projection Data. Int. J. Biomed. Imaging 2008, Article ID 427989, 6p (2008). doi:10.1155/2008/427989

105. Yu, H., Wang, G.: Compressed sensing based interior tomography. Phys. Med. Biol. **54**, 2791–2805 (2009)
106. Zalcman, L.: Offbeat integral geometry. Am. Math. Mon. **87**, 161–175 (1980)
107. Zalcman, L.: Uniqueness and nonuniqueness for the Radon transform. Bull. Lond. Math. Soc. **14**(3), 241–245 (1982). doi:10.1112/blms/14.3.241
108. Zampighi, G., Zampighi, L., Fain, N., Wright, E., Cantelle, F., Lanzavecchia, S.: Conical tomography II: a method for the study of cellular organelles in thin sections. J. Struct. Biol. **151**(3), 263–274 (2005)

Mathematical Methods in PET and SPECT Imaging

Athanasios S. Fokas and George A. Kastis

Contents

1 Introduction 904
2 Background 907
The Importance of PET and SPECT 907
The Mathematical Foundation of the IART 910
A General Methodology for Constructing Transform Pairs 911
3 The Inverse Radon Transform and the Inverse Attenuated Radon Transform 912
The Construction of the Inverse Radon Transform 913
The Construction of Inverse Attenuated Radon Transform 917
4 SRT for PET 918
Comparison between FBP and SRT for PET 922
5 SRT for SPECT 930
6 Conclusion 933
7 Cross-References 934
References 934

Abstract

In this chapter, we present the mathematical formulation of the inverse Radon transform and of the inverse attenuated Radon transform (IART), which are used in PET and SPECT image reconstruction, respectively. Using a new method for deriving transform pairs in one and two dimensions, we derive the inverse Radon transform and the IART. Furthermore, we discuss an alternative approach for

A.S. Fokas (✉)
University of Technology Darmstadt, Darmstadt, Germany
e-mail: t.fokas@damtp.cam.ac.uk

G.A. Kastis
Research Center of Mathematics, Academy of Athens, Athens, Greece
e-mail: gkastis@academyofathens.gr

© Springer Science+Business Media New York 2015

O. Scherzer (ed.), *Handbook of Mathematical Methods in Imaging*,
DOI 10.1007/978-1-4939-0790-8_45

computing the Hilbert transform using cubic splines. This new approach, which is referred to as spline reconstruction technique, is formulated in the physical space, in contrast to the well-known filtered backprojection (FBP) algorithm which is formulated in the Fourier space. Finally, we present the results of several rigorous studies comparing FBP with SRT for PET. These studies, which use both simulated and real data and which employ a variety of image quality measures including contrast and bias, indicate that SRT has certain advantages in comparison with FBP.

1 Introduction

Positron emission tomography (PET) is an important, noninvasive, nuclear medicine modality that measures the in vivo distribution of imaging agents labeled with positron-emitting radionuclides. The importance of PET in detecting, staging, and monitoring the progress of several diseases has been established in a plethora of rigorous clinical studies [1]. Furthermore, small-animal PET is becoming an essential imaging modality for preclinical research [2–4], as well as for drug development and discovery [5].

Single-photon emission computed tomography (SPECT) is the most widely available diagnostic imaging technique which uses short-lived radiographic isotopes. In particular, it is extensively used in cardiology. For example, in 2012, nine million myocardial perfusion scintigraphic imaging studies were performed in the USA.

Image reconstruction is an essential component in tomographic medical imaging, allowing a tomographic image to be obtained from a set of two-dimensional projections. The existing image reconstruction methods can be classified into two broad categories: (a) analytic methods and (b) iterative (or algebraic) methods. In what follows, we will discuss only analytic methods. In this respect, we first consider a simple model for SPECT: let $\mu(x_1, x_2)$ denote the gamma-ray attenuation coefficient of the tissue at the point (x_1, x_2). This means that a gamma ray traveling a small distance $\Delta\tau$ at (x_1, x_2) suffers a relative intensity loss

$$\frac{\Delta I}{I} = -\mu \Delta\tau. \tag{1}$$

We denote by I_o the initial intensity of a gamma ray and by I_f the measured intensity after undergoing attenuation through tissue of length L; the above equation can be written in the form

$$I_f = I_o \exp\left\{-\int_L \mu(\tau)\, d\tau\right\}. \tag{2}$$

This equation expresses the well-known Beer's law.

Let $f(x_1, x_2)$ denote the distribution in the tissue of the radioactive material under consideration, and let $L(x)$ denote the part of the ray from the tissue to the detector. Then, in SPECT, it is assumed that the following integral, I_{SPECT}, is known from the measurements:

$$I_{\text{SPECT}} = \int_L \exp\left\{-\int_{L(x)} \mu \, ds\right\} f \, d\tau, \tag{3}$$

where the integral is measured over a finite number of lines L. Equation (3) is valid under the following assumptions:

(a) The lines L are lying within a specified imaging plane,
(b) The imaging system has perfect imaging characteristics,
(c) The detector is collimated and can pick up radiation only along the straight lines L,
(d) The collimator has an infinitely high spatial resolution,
(e) The attenuation coefficient μ obeys Eq. (2),
(f) There is no scatter radiation in the object.

In PET, the integral in Eq. (3) simplifies to

$$I_{\text{PET}} = \exp\{-\hat{\mu}\} \int_L f \, d\tau, \tag{4}$$

where $\hat{\mu}$ denotes the *Radon transform* of μ, i.e.,

$$\hat{\mu} = \int_L \mu \, d\tau. \tag{5}$$

In order to derive Eq. (4) from Eq. (3), we recall that positrons eject particles pairwise in opposite directions, and in PET the radiation in opposite directions is measured simultaneously. Thus,

$$\int_{L(x)} \mu \, ds = \int_{L_-(x)} \mu \, ds + \int_{L_+(x)} \mu \, ds = \int_L \mu \, ds \tag{6}$$

and Eq. (3) becomes Eq. (4).

In both PET and SPECT, a projection is formed by combining a set of line integrals along a line L. Here, we will consider parallel-ray integrals at a specific angle θ, i.e., we will assume a parallel-beam geometry.

We recall that the Radon transform of the X-ray attenuation coefficient denoted by μ_{CT} is measured via computed tomography (CT). Using the values of μ_{CT}, it is possible to estimate $\mu(x_1, x_2)$ by appropriate scaling (this is necessary due to the energy difference between X-rays and gamma rays).

Equation (4) implies that in PET one needs to reconstruct a function f from the knowledge of its Radon transform denoted by $\hat{f}$,

$$\hat{f} = \int_L f \, d\tau, \tag{7}$$

where $\hat{f}$ can be computed via

$$\hat{f} = I_{\text{PET}} \exp\{\hat{\mu}\}. \tag{8}$$

Hence, both CT and PET involve the computation of a function from the knowledge of its Radon transform. More specifically, in CT one needs to compute μ_{CT} from the knowledge of $\hat{\mu}_{\text{CT}}$, whereas in PET one needs to compute f from the knowledge of $I_{\text{PET}} \exp\{\hat{\mu}\}$. The relevant formula, known as the *inverse Radon transform*, is given by

$$f(x_1, x_2) = \frac{1}{4\pi}(\partial_{x_1} - i\,\partial_{x_2}) \int_0^{2\pi} e^{i\theta} J(x_1, x_2, \theta)\, d\theta, \tag{9}$$

where J is defined in terms of $\hat{f}$ by

$$J(x_1, x_2, \theta) = \frac{1}{i\pi} \oint_L \frac{\hat{f}(\rho, \theta)\, d\rho}{\rho - (x_2 \cos\theta - x_1 \sin\theta)}, \quad 0 \le \theta < 2\pi, \tag{10}$$

and throughout this chapter, $\oint$ denotes the principal value integral.

Let $\hat{f}_\mu$ denote the right-hand side (RHS) of Eq. (3). We call $\hat{f}_\mu$ the *attenuated Radon transform* of f (with attenuation specified by the given function μ). Hence, SPECT involves the computation of a function f from the knowledge of its attenuated Radon transform $\hat{f}_\mu$ and of the function μ. The relevant formula is known as the *inverse attenuated Radon transform* (IART).

According to Eq. (9), the numerical implementation of the inverse Radon transform involves the computation of the Hilbert transform $(Hf)(\rho)$ of a given function $f(\rho)$, where

$$(Hf)(\rho) = \frac{1}{\pi} \oint_L \frac{f(r)}{r - \rho}\, dr. \tag{11}$$

The most well-known method for this computation uses the fast Fourier transform technique, exploiting the fact that the Hilbert transform is a convolution. The application of this method to the computation of the inverse Radon transform is called *filtered backprojection* (FBP).

Most PET systems have options for both FBP and Ordered Subset Expectation Maximization (OSEM), which is the predominant iterative technique.

The analytical approach to SPECT requires the numerical implementation of the IART. However, since no analytical formula was available until recently for the IART, currently the FBP implementation to SPECT generally makes the crude approximation of $\mu = 0$ or incorporates a uniform attenuation map [6]. On the other hand, since OSEM is based on an iterative statistical approach, it does *not* require the analytical inversion formula, and hence OSEM considers the actual values of μ. Hence, in SPECT, OSEM produces more accurate images than FBP. However, for several reasons including speed (1 s vs. 2.5 min for a typical study), FBP is still used extensively. Indeed, most SPECT systems have options for both FBP and OSEM, and both are used clinically. Actually, some clinicians prefer FBP for quantification.

In this review, we will discuss an alternative approach for computing the Hilbert transform. This approach, in contrast to FBP which is formulated in the Fourier space, is formulated in the physical space, and it is based on "custom-made" cubic splines. We will refer to the application of this approach to the computation of the inverse Radon transform and to the IART as *spline reconstruction technique* (SRT) for PET and SRT for SPECT, respectively.

The results of rigorous studies comparing FBP with SRT for PET are published in [7]. In these studies, which use both simulated and real data, by employing a variety of image quality measures, including contrast and bias, it is established that SRT has certain advantages in comparison with FBP. As a result of these studies, SRT is now included in STIR (Software for Tomographic Image Reconstruction) [8], which is a widely used open-source software library in tomographic imaging (FBP and SRT are the only algorithms based on analytical formulae which are incorporated in STIR).

Rigorous studies comparing SRT for SPECT with FBP and OSEM are work in progress. Preliminary results suggest that SRT is preferable to FBP and also that SRT is comparable with OSEM; see Fig. 1 [9].

The clinical importance of PET and SPECT, as well as the mathematical formulation of the inverse Radon transform and of the IART, is further discussed in the following section. A new method for deriving transforms in one and two dimensions was introduced in [10]. Using this method, both the inverse Radon transform and the IART are derived in Sect. 3. The SRT for PET is presented in Sect. 4, and the studies of Kastis et al. [7] comparing SRT with FBP for PET are reviewed in Sect. 5. SRT for SPECT is briefly discussed in Sect. 6.

2 Background

The Importance of PET and SPECT

In 1964, the research group of Dr. David E. Kuhl, known as the "Father of Emission Tomography," developed the Mark II SPECT series, which is a single-emission computed tomography camera. Using this unit, this group succeeded in producing

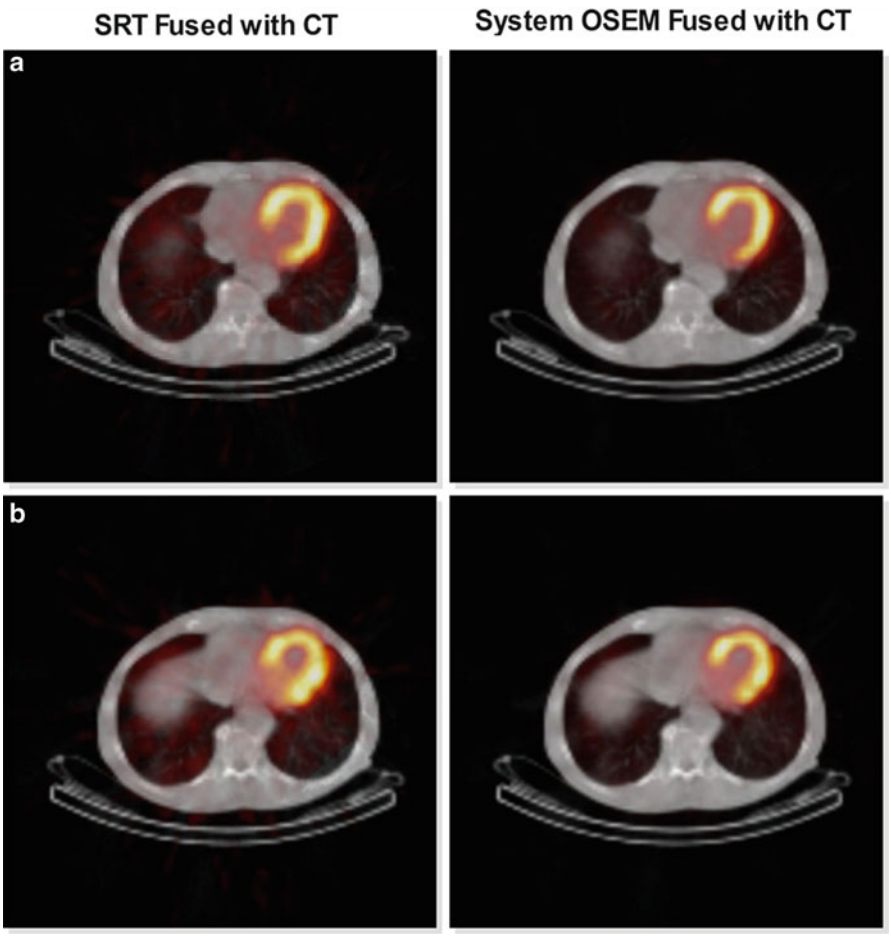


Fig. 1 Reconstructions of two consecutive slices (**a** and **b**) of the myocardium acquired with SPECT fused with CT scan. The images on the *right* are obtained by the currently used commercial software and on the *left* with SRT

the world's first tomographic images of the human body. This achievement was considerably earlier than Godfrey Hounsfield's development of the X-ray CT device in 1972. A crucial element in the evolution of emission tomography was the choice of the radioisotope injected in the patient. An important such class is those radioisotopes that emit positrons. An emitted positron almost immediately combines with a nearby electron; they annihilate each other, emitting in the process two gamma rays traveling in nearly opposite directions. PET cameras detect these gamma rays.

FDG (^{18}F-2-deoxy-2-fluoro-D-glucose) is the most suitable positron emitter that can be used with humans. It is a deoxyglucose analog with the normal hydroxyl

group in the glucose molecule substituted with the radioactive isotope ^{18}F. FDG is taken up by high-glucose-using cells, and it is metabolically trapped following phosphorylation by the hexokinase enzyme [11]. Thus, the measurement of the positron-emitting ^{18}F provides a quantitative measure of the glucose consumed in the corresponding tissue. Glucose metabolism in cancerous tissue is higher than in normal tissue. Thus, PET-CT devices are proving to be highly effective in the early detection of cancer. For example, among 91 patients with non-small-cell lung cancer, 38 pathologically confirmed mediastinal lymph node metastases were missed by CT, whereas among 98 similar patients, only 21 metastases were missed by PET-CT [12] (this means that in the former case, there were 38 futile thoracotomies, whereas in the latter case, there were only 21). Furthermore, using more specific positron emitters, it is possible to obtain additional useful information. For example, using F-fluoro-17-β-estradiol, it is possible to access in vivo the density of estrogen receptors as well as to monitor the response of the treatment for estrogen receptor-positive breast cancer. Similarly, using ^{18}F-annexin V, it is possible to follow tumor cell apoptosis (death) in vivo, as well as to monitor treatment response to various cancer types. It should also be noted that PET-CT is useful for prognosis. For example, the 2-year failure-free survival rate of patients with positive scans after chemotherapy treatment for early-stage Hodgkin's lymphoma was 69 %, as compared with 95 % of patients with negative scans [13].

In addition to oncology, PET is now used in several other areas of medicine, including cardiology and neurology. As an example of the usefulness of PET in neurology, we note that PET with the use of the Pittsburgh compound B(PIB) can be used to quantify the concentration of amyloid-beta (Aβ) deposition in the brain, a precursor of Alzheimer's disease [14].

SPECT is used extensively in cardiology, oncology, and neurology. In cardiology, SPECT is part of the common stress-testing procedures for the evaluation of chest pain [15]. As an example of the use of SPECT in oncology, we mention the study in [16], where lymphoscintigraphy was performed using SPECT-CT to establish a preoperative road map of lymph nodes that are at risk for metastatic melanoma and to facilitate intraoperative identification of the "sentinel" nodes (the rationale for sentinel-node biopsy relies on the concept that different regions of the skin have specific patterns of lymphatic drainage to the regional nodes, and for a given region of the skin, there exists at least a specific node, the sentinel node, in which lymphatic vessels drain first, and it is this node the most likely first site of metastasis).

In neurology, SPECT can be used to image the distribution of the blood flow in the brain, and this has been used in a plethora of pathological situations. Furthermore, by employing specific radioisotopes, it is possible to obtain useful information about several diseases. For example, using SPECT with ^{123}I-labeled CIT, it is shown in [17] that although levodopa treatment is highly effective as dopamine replacement in Parkinson's disease, this treatment apparently downregulates the endogenous dopamine transporters.

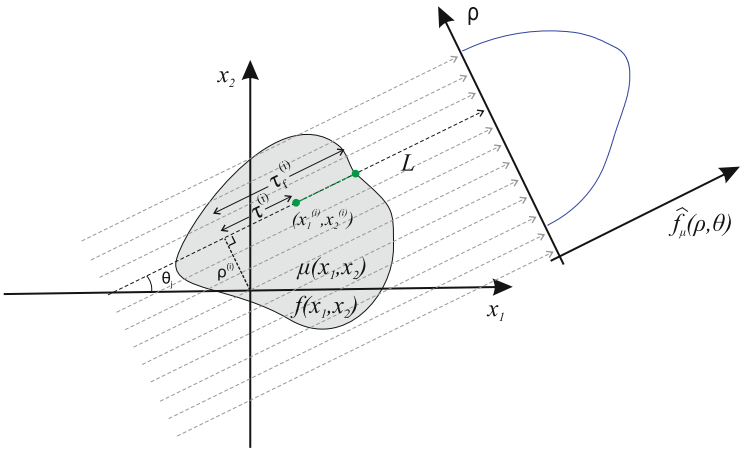


Fig. 2 Parallel-beam projections through an object with attenuation coefficient $\mu(x_1, x_2)$. The local and Cartesian coordinates of the mathematical formulation are indicated

The Mathematical Foundation of the IART

Consider a line L specified by two real numbers ρ and θ, where $-\infty < \rho < \infty$ and $0 \leq \theta < 2\pi$ (Fig. 2). A unit vector along this line is given by

$$\mathbf{e}^{\parallel} = (\cos\theta, \sin\theta)\,. \tag{12}$$

A unit vector perpendicular to L is given by

$$\mathbf{e}^{\perp} = (-\sin\theta, \cos\theta)\,. \tag{13}$$

A point $\mathbf{x}$ on this line can be expressed in the form

$$\mathbf{x} = \rho\,\mathbf{e}^{\perp} + \tau\,\mathbf{e}^{\parallel}. \tag{14}$$

Hence,

$$x_1 = \tau\cos\theta - \rho\sin\theta, \tag{15}$$

$$x_2 = \tau\sin\theta + \rho\cos\theta. \tag{16}$$

Solving these equations for (τ, ρ) in terms of (x_1, x_2, θ), we find

$$\tau = x_2\sin\theta + x_1\cos\theta, \tag{17}$$

$$\rho = x_2\cos\theta - x_1\sin\theta. \tag{18}$$

Writing (x_1, x_2) in terms of the local coordinates (ρ, θ), it follows that the equation defining the Radon transform, $\hat{f}$, of a function f takes the form

$$\hat{f}(\rho,\theta) = \int_{-\infty}^{\infty} f\,(\tau\cos\theta - \rho\sin\theta, \tau\sin\theta + \rho\cos\theta)\,d\tau, \\ -\infty < \rho < \infty,\ \ 0 \le \theta < 2\pi. \tag{19}$$

Similarly, the following equation provides the definition of the attenuated Radon transform, $\hat{f}_\mu$ (with attenuation specified by μ), of a function f:

$$\hat{f}_\mu(\rho,\theta) = \int_{-\infty}^{\infty} \exp\left\{-\int_{\tau}^{\infty} \mu(s\cos\theta - \rho\sin\theta, s\sin\theta + \rho\cos\theta)ds\right\} \times \\ f(\tau\cos\theta - \rho\sin\theta, \tau\sin\theta + \rho\cos\theta)d\tau, \quad -\infty < \rho < \infty, \quad 0 \le \theta < 2\pi. \tag{20}$$

The inverse Radon transform, expressing f in terms of $\hat{f}$, is given by Eqs. (9) and (10). Similarly, the IART, expressing f in terms of $\hat{f}_\mu$ and μ, is given by

$$f(x_1,x_2) = \frac{1}{4\pi}(\partial_{x_1} - i\,\partial_{x_2}) \int_0^{2\pi} e^{i\theta} J(x_1,x_2,\theta)d\theta, \qquad -\infty < x_1, x_2 < \infty, \tag{21}$$

where the function J is defined by

$$J(x_1,x_2,\theta) = \exp\left\{\int_{\tau}^{\infty} \mu(s\cos\theta - \rho\sin\theta, s\sin\theta + \rho\cos\theta)ds\right\} \\ \times L_\mu(\rho,\theta)\hat{f}_\mu(\rho,\theta)\Big|_{\substack{\tau = x_2\sin\theta + x_1\cos\theta \\ \rho = x_2\cos\theta - x_1\sin\theta}} \tag{22}$$

and the operator $L_\mu(\rho,\theta)$ is defined by

$$L_\mu(\rho,\theta) = \exp\{P^-\hat{\mu}(\rho,\theta)\}\,P^-\exp\{-P^-\hat{\mu}(\rho,\theta)\} \\ + \exp\left\{-P^+\hat{\mu}(\rho,\theta)\right\}P^+\exp\left\{P^+\hat{\mu}(\rho,\theta)\right\}, \tag{23}$$

with $\hat{\mu}(\rho,\theta)$ denoting the Radon transform of $\mu(x_1,x_2)$ and the operators $P^\pm$ denoting the usual projection operators in the variable ρ, i.e.,

$$(P^\pm g)(\rho) = \pm\frac{g(\rho)}{2} + \frac{1}{2i\pi}\oint_{-\infty}^{\infty}\frac{g(r)}{r-\rho}dr, \qquad -\infty < \rho < \infty. \tag{24}$$

A General Methodology for Constructing Transform Pairs

The study of nonlinear integrable equations led to the emergence of a general method for deriving transform pair in one and two dimensions [10]. In particular,

using this method, it was shown in [10] that the two-dimensional Fourier transform of a function $u(x_1, x_2)$ can be constructed via the *spectral analysis* of the equation

$$\frac{\partial \Psi}{\partial x_1} + \frac{i\,\partial \Psi}{\partial x_2} - k\Psi = u(x_1, x_2),\ (x_1, x_2) \in \mathbf{R}^2,\ k \in \mathbf{C}, \tag{25}$$

where $\Psi(x_1, x_2, k, \bar{k})$ is a scalar function and bar denotes complex conjugate.

R. Novikov and one of the authors re-derived in [18] the Radon transform by performing the spectral analysis of the equation

$$\frac{1}{2}\left(k + \frac{1}{k}\right)\frac{\partial \Psi}{\partial x_1} + \frac{1}{2i}\left(k - \frac{1}{k}\right)\frac{\partial \Psi}{\partial x_2} = f(x_1, x_2),\ (x_1, x_2) \in \mathbf{R}^2,\ k \in \mathbf{C}. \tag{26}$$

Although the Radon transform can be derived in a much simpler way by using the two-dimensional Fourier transform, the advantage of the derivation of [18] was demonstrated later by Novikov [19] who showed that the inverse attenuated Radon transform can be derived by applying a similar analysis to the following slight generalization of Eq. (26):

$$\frac{1}{2}\left(k + \frac{1}{k}\right)\frac{\partial \Psi}{\partial x_1} + \frac{1}{2i}\left(k - \frac{1}{k}\right)\frac{\partial \Psi}{\partial x_2} - \mu(x_1, x_2)\Psi = f(x_1, x_2), (x_1, x_2) \in \mathbf{R}^2,\ k \in \mathbf{C}. \tag{27}$$

3 The Inverse Radon Transform and the Inverse Attenuated Radon Transform

In what follows, we discuss an algorithmic approach for the construction of a transform pair $\{f, \hat{f}\}$. The relevant analysis, which is usually referred to as *spectral analysis*, involves two main steps: (i) solve a given eigenvalue equation in terms of f. If k denotes the eigenvalue parameter, this involves constructing a solution Ψ of the given eigenvalue equation which is *bounded* for all complex values of k. This problem will be referred to as the *direct problem*. (ii) Using the fact that Ψ is bounded for all complex k, construct an alternative representation of Ψ which (instead of depending on f) depends on some "spectral function" of f denoted by $\hat{f}$. This problem will be referred to as the *inverse problem*.

It turns out that the inverse problem gives rise to certain problems in complex analysis known as the Riemann-Hilbert and the d-bar problems. Indeed, for certain eigenvalue problems, the function Ψ is *sectionally analytic* in k, i.e., it has different representations in different domains of the complex k-plane, and each of these representations is analytic. In this case, if the "jumps" of these representations across the different domains can be expressed in terms of $\hat{f}$, then it is possible to reconstruct Ψ as the solution of a Riemann-Hilbert problem which is uniquely defined in terms of $\hat{f}$. However, for a large class of eigenvalue problems, there

exists a domain in the complex k-plane where Ψ is *not* analytic. In this case, if $\partial\Psi/\partial\bar{k}$ can be expressed in terms of $\hat{f}$, then Ψ can be reconstructed through the solution of a d-bar problem which is uniquely defined in terms of $\hat{f}$.

We recall that the classical derivation of transform pairs involves the integration in the complex k-plane of an appropriate Green's function. However, this derivation is based on the assumption that the Green's function is an analytic function of k and it also assumes completeness. The assumption of analyticity corresponds to the case that Ψ is sectionally analytic. Therefore, the approach reviewed here has the advantage that not only it provides a simpler approach to deriving classical transforms avoiding the problem of completeness, but also it can be applied to problems that the associated Green's function is *not* an analytic function of k.

In addition to the construction of the Inverse attenuated Radon transform, the approach reviewed here has led to the construction of the X-ray fluorescence computed tomography [20].

The Construction of the Inverse Radon Transform

Let $S(\mathbf{R}^2)$ denote the space of Schwartz functions. Define the Radon transform $\hat{f}(\rho,\theta)$ of the function $f(x_1,x_2) \in S(\mathbf{R}^2)$ by Eq. (19). Then, for all $(x_1,x_2) \in \mathbf{R}^2$, $f(x_1,x_2)$ is given by Eq. (9).

We will derive the Radon transform pair by performing the spectral analysis of the eigenvalue Eq. (26). In order to solve the direct problem, we first simplify Eq. (26) by introducing a change of variables from (x_1,x_2) to $(z,\bar{z})$, where z is defined by

$$z = \frac{1}{2i}\left(k - \frac{1}{k}\right)x_1 - \frac{1}{2}\left(k + \frac{1}{k}\right)x_2, \tag{28}$$

and $\bar{z}$ follows via complex conjugation, i.e.,

$$\bar{z} = -\frac{1}{2i}\left(\bar{k} - \frac{1}{\bar{k}}\right)x_1 - \frac{1}{2}\left(\bar{k} + \frac{1}{\bar{k}}\right)x_2, \quad (x_1,x_2) \in \mathbf{R}^2, \quad k \in \mathbf{C}.$$

Using the identities

$$\partial_{x_1} = \frac{1}{2i}\left(k - \frac{1}{k}\right)\partial_z - \frac{1}{2i}\left(\bar{k} - \frac{1}{\bar{k}}\right)\partial_{\bar{z}} \tag{29}$$

and

$$\partial_{x_2} = -\frac{1}{2}\left(k + \frac{1}{k}\right)\partial_z - \frac{1}{2}\left(\bar{k} + \frac{1}{\bar{k}}\right)\partial_{\bar{z}}, \tag{30}$$

Eq. (26) becomes

$$\nu(|k|)\frac{\partial \Psi}{\partial \bar{z}}(x_1, x_2, k) = f(x_1, x_2), \quad (x_1, x_2) \in \mathbf{R}^2, \quad k \in \mathbf{C}, \tag{31}$$

where the function $\nu(|k|)$ is defined by

$$\nu(|k|) = \frac{1}{2i}\left(\frac{1}{|k|^2} - |k|^2\right). \tag{32}$$

We supplement Eq. (31) with the boundary condition

$$\Psi = O\left(\frac{1}{z}\right), \quad z \to \infty. \tag{33}$$

In order to solve Eqs. (31) and (33), we will use the Pompeiu formula, namely,

$$f(z, \bar{z}) = \frac{1}{2i\pi}\int_{\partial D}\frac{f(\zeta, \bar{\zeta})d\zeta}{\zeta - z} + \frac{1}{2i\pi}\int\int_D \frac{\partial f}{\partial \bar{\zeta}}(\zeta, \bar{\zeta})\frac{d\zeta \wedge d\bar{\zeta}}{\zeta - z}, \quad z \in D, \tag{34}$$

where ∂D denotes the boundary of the domain D. Using this formula, we find

$$\Psi = \frac{1}{2\pi i}\int\int_{\mathbf{R}^2}\frac{f(x_1', x_2')}{\nu(|k|)}\frac{dz' \wedge d\bar{z}'}{z' - z}, \ (x_1, x_2) \in \mathbf{R}^2, \quad k \in \mathbf{C}, \quad |k| \neq 1. \tag{35}$$

Hence, using the identity

$$dz \wedge d\bar{z} = \frac{1}{2i}\left||k|^2 - \frac{1}{|k|^2}\right| dx_1 dx_2 \tag{36}$$

it follows that for all $(x_1, x_2) \in \mathbf{R}^2$ and $k \in \mathbf{C}$, $|k| \neq 1$, Ψ satisfies

$$\Psi(x_1, x_2, k) = \frac{1}{2\pi i}\text{sgn}\left(\frac{1}{|k|^2} - |k|^2\right)\int\int_{\mathbf{R}^2} f(x_1', x_2')\frac{dx_1' dx_2'}{z' - z}. \tag{37}$$

If k is either inside or outside the unit circle, the only dependence of Ψ on k is through z' and z; thus, Ψ is a sectionally analytic function with a "jump" across the unit circle of the complex k-plane. Equation (37) provides the solution of the direct problem.

In order to solve the inverse problem, we will formulate a Riemann-Hilbert problem in the complex k-plane. In this respect, we note that Eq. (37) implies

$$\Psi = O\left(\frac{1}{k}\right), \quad k \to \infty. \tag{38}$$

Furthermore, we will show that for all $(x_1, x_2) \in \mathbf{R}^2$, Ψ satisfies the following "jump" condition:

$$\Psi^+ - \Psi^- = i(H\hat{f})(\rho,\theta), \quad -\infty < \rho < \infty, \quad 0 \le \theta < 2\pi, \tag{39}$$

where H denotes the Hilbert transform in the variance ρ defined in Eq. (11). This equation is a direct consequence of the following equations: let Ψ^+ and Ψ^- denote the limits of Ψ as k approaches the unit circle from inside and outside, i.e.,

$$\Psi^{\pm} \equiv \lim_{\varepsilon \to 0} \Psi\left(x_1, x_2, (1 \mp \varepsilon)e^{i\theta}\right). \tag{40}$$

Then, for all $(x_1, x_2) \in \mathbf{R}^2$,

$$\Psi^{\pm} = \mp(P^{\mp}\hat{f})(\rho,\theta) - \int_{\tau}^{\infty} F(\rho, s, \theta)ds, \quad -\infty < \rho < \infty, \quad 0 \le \theta < 2\pi, \tag{41}$$

where $P^{\pm}$ denote the usual projectors in the variable ρ, i.e.,

$$(P^{\pm}\hat{f})(\rho) = \pm\frac{f(\rho)}{2} + \frac{1}{2i}(Hf)(\rho) \tag{42}$$

and F denotes f in the coordinates (ρ, τ, θ), i.e.,

$$F(\rho, \tau, \theta) = f(\tau\cos\theta - \rho\sin\theta, \tau\sin\theta + \rho\cos\theta). \tag{43}$$

Indeed, in order to derive Eq. (41), we note that the definition of z implies

$$(z - z') = \frac{1}{2i}\left(k - \frac{1}{k}\right)(x_1 - x_1') - \frac{1}{2}\left(k + \frac{1}{k}\right)(x_2 - x_2'). \tag{44}$$

Let

$$k^+ = (1-\varepsilon)e^{i\theta}, \quad k^- = (1+\varepsilon)e^{i\theta}, \quad 0 \le \theta < 2\pi, \quad \varepsilon > 0. \tag{45}$$

Thus

$$\left(k^+ \mp \frac{1}{k^+}\right) = (1-\varepsilon)e^{i\theta} \mp (1+\varepsilon)e^{-i\theta} + O(\varepsilon^2) \tag{46}$$

and similarly for $(k^- \mp 1/k^-)$.

Hence, for computing $\Psi^{\pm}$ via Eq. (37), we must use the formulae

$$\begin{aligned} z' - z &= \left(x_1' - x_1\right)\sin\theta - \left(x_2' - x_2\right)\cos\theta \\ \pm i\varepsilon &\left[\left(x_1' - x_1\right)\cos\theta + \left(x_2' - x_2\right)\sin\theta\right] + O(\varepsilon^2). \end{aligned} \tag{47}$$

We recall that solving Eqs. (15) and (16) for (ρ, τ) in terms of (x_1, x_2) we find Eqs. (17) and (18). The Jacobian of this transformation equals 1, hence $dx_1 dx_2 =$

$d\rho d\tau$. Replacing in Eq. (37) $z - z'$ by the RHS of equation (47) and then changing variables in the resulting equation from (x_1', x_2') to (ρ', τ') and from (x_1, x_2) to (ρ, τ), we find

$$\Psi^{\pm} = \mp \frac{1}{2i\pi} \lim_{\varepsilon \to 0} \int \int_{\mathbf{R}^2} \frac{F(\rho', \tau', \theta) d\rho' d\tau'}{\rho' - [\rho \pm i\varepsilon(\tau' - \tau)]}. \tag{48}$$

In order to evaluate this limit, we must control the sign of $\tau' - \tau$. This suggests splitting the integral over $d\tau'$ as shown below:

$$\Psi^{\pm} = \mp \frac{1}{2i\pi} \lim_{\varepsilon \to 0} \int_{-\infty}^{\infty} \left\{ \int_{-\infty}^{\tau} \frac{F d\tau'}{\rho' - [\rho \pm i\varepsilon(\tau' - \tau)]} + \int_{\tau}^{\infty} \frac{F d\tau'}{\rho' - [\rho \pm i\varepsilon(\tau' - \tau)]} \right\} d\rho'. \tag{49}$$

In the first and second integral above, $\tau' - \tau$ is negative and positive, respectively: hence

$$\begin{aligned} \Psi^{\pm} = & \mp \frac{1}{2i} \int_{-\infty}^{\tau} \left\{ \mp iF(\rho, \tau', \theta) + (HF)(\rho, \tau', \theta) \right\} d\tau' \\ & \mp \frac{1}{2i} \int_{\tau}^{\infty} \left\{ \pm iF(\rho, \tau', \theta) + (HF)(\rho, \tau', \theta) \right\} d\tau'. \end{aligned} \tag{50}$$

Adding and subtracting in the RHS of this equation the expression

$$\mp \frac{1}{2i} \int_{\tau}^{\infty} (\mp) iF(\rho, \tau', \theta) d\tau', \tag{51}$$

we find Eqs. (41).

The sectionally analytic function Ψ satisfies the estimate (38) and has a jump across the unit circle; thus, for all $(x_1, x_2) \in \mathbf{R}^2$, it admits the following representation:

$$\Psi = \frac{1}{2i\pi} \int_0^{2\pi} \frac{(\Psi^+ - \Psi^-)(\rho, \theta') i e^{i\theta'} d\theta'}{e^{i\theta'} - k}, \quad k \in \mathbf{C}, \quad |k| \neq 1, \quad -\infty < \rho < \infty. \tag{52}$$

Replacing in this equation $\Psi^+ - \Psi^-$ by the RHS of (39), we find the following expression valid for all $(x_1, x_2) \in \mathbf{R}^2$:

$$\Psi = -\frac{1}{2i\pi} \int_0^{2\pi} \frac{e^{i\theta'} (H\hat{f})(\rho, \theta') d\theta'}{e^{i\theta'} - k}, \quad k \in \mathbf{C}, \quad |k| \neq 1, \quad -\infty < \rho < \infty. \tag{53}$$

This expression provides the solution of the inverse problem.

Using Eqs. (26) and (53), it is straightforward to express f in terms of $\hat{f}$. One way of achieving this is to replace in Eq. (26) Ψ by the RHS of Eq. (53). A simpler alternative way is to compute the large k behavior of Ψ: Equation (53) implies

$$\Psi \sim \left\{ \frac{1}{2i\pi} \int_0^{2\pi} e^{i\theta} (H\hat{f})(\rho,\theta) d\theta \right\} \frac{1}{k} + O\left(\frac{1}{k^2}\right), \quad k \to \infty. \tag{54}$$

Substituting this expression in Eq. (26), we find that the $O(1)$ term of Eq. (26) yields

$$f = \frac{1}{4i\pi} \left(\partial_{x_1} + \frac{1}{i} \partial_{x_2} \right) \int_0^{2\pi} e^{i\theta} (H\hat{f})(\rho,\theta) d\theta, \tag{55}$$

which is Eq. (9).

The Construction of Inverse Attenuated Radon Transform

Let $k^{\pm}$ denote the limiting values of $k \in \mathbf{C}$ as it approaches the unit circle in the complex k-plane from inside and outside the unit circle; see Eq. (45). Let z be defined in terms of $(x_1, x_2) \in \mathbf{R}^2$ and $k \in \mathbf{C}$ by Eq. (28), and let $\nu(|k|)$ be defined by Eq. (32). Then

$$\lim_{k \to k^{\pm}} \left\{ \partial_{\bar{z}}^{-1} \left(\frac{f(x_1, x_2)}{\nu(|k|)} \right) \right\} = \mp (P^{\mp} \hat{f})(\rho,\theta) - \int_{\tau}^{\infty} F(\rho, s, \theta) ds, \quad -\infty < \rho < \infty, \quad 0 \le \theta < 2\pi, \tag{56}$$

where $\hat{f}$ is the Radon transform of f (see Eq. (19)), $P^{\pm}$ are the usual projectors in the variable ρ (see Eq. (42)), (ρ, τ) are defined in terms of (x_1, x_2) by Eqs. (17) and (18), and F denotes f in the variables (ρ, τ, θ) (see Eq. (43)). Indeed, Eq. (56) is direct consequence of Eqs. (31) and (41).

It turns out that the derivation of the attenuated Radon transform pair is a direct consequence of Eqs. (27) and (56). Indeed, define the attenuated Radon transform $\hat{f}_{\mu}(\rho,\theta)$ of the function $f(x_1, x_2) \in S(\mathbf{R}^2)$ by Eq. (20). Then, $f(x_1, x_2)$ is given by Eq. (21).

In order to derive the above result, we note that Eq. (27) can be rewritten in the form

$$\frac{\partial \Psi}{\partial \bar{z}} + \frac{\mu}{\nu} \Psi = \frac{f}{\nu}. \tag{57}$$

Hence

$$\frac{\partial}{\partial \bar{z}} \left[\Psi e^{\partial_{\bar{z}}^{-1}(\frac{\mu}{\nu})} \right] = \frac{f}{\nu} e^{\partial_{\bar{z}}^{-1}(\frac{\mu}{\nu})}, \tag{58}$$

or

$$\Psi e^{\partial_{\bar{z}}^{-1}(\frac{\mu}{\nu})} = \partial_{\bar{z}}^{-1} \left[\left(\frac{f}{\nu} \right) e^{\partial_{\bar{z}}^{-1}(\frac{\mu}{\nu})} \right], \quad (x_1, x_2) \in \mathbf{R}^2, \quad k \in \mathbf{C}. \tag{59}$$

This equation provides the solution of the direct problem, i.e., it defines a sectionally analytic function Ψ with the estimate (38), which has a jump across the unit circle of the complex k-plane. Hence, Ψ is given by Eq. (52). All that remains is to determine the jump $\Psi^+ - \Psi^-$. This involves computing the limits as $k \to k^\pm$ of $\partial_{\bar{z}}^{-1}(\mu/\nu)$; thus, it can be achieved using Eq.(56). Let M denote μ in the variables (ρ, τ, θ): then using in Eq. (59) the identity (56) with f replaced by μ, we find

$$\Psi^{\pm} e^{\mp P^{\mp}\hat{\mu}} e^{-\int_\tau^\infty M(\rho,s,\theta)ds} = \lim_{k\to k^\pm} \partial_{\bar{z}}^{-1} \left\{ \frac{f}{\nu} e^{\mp P^{\mp}\hat{\mu}} e^{-\int_\tau^\infty M(\rho,s,\theta)ds} \right\}. \tag{60}$$

For the computation of the RHS of this equation, we use again Eq. (56) where f is now replaced by f times the two exponentials appearing in the curly bracket of Eq. (60). Hence, the RHS of Eq. (60) yields

$$\mp P^{\mp} e^{\mp P^{\mp}\hat{\mu}} \hat{f}_\mu - \int_\tau^\infty F(\rho, \tau', \theta) e^{\mp P^{\mp}\hat{\mu}} e^{-\int_{\tau'}^\infty M(\rho,s,\theta)ds} d\tau'. \tag{61}$$

The term $\exp[\mp P^{\mp}\hat{\mu}]$ is independent of τ'; hence, this term comes out of the integral $\int_\tau^\infty$, and furthermore, the same terms appear in the left-hand side of Eq. (60). Hence,

$$\Psi^+ - \Psi^- = -J, \tag{62}$$

where J is defined in Eq. (22). Then Eq. (52) yields

$$\Psi = -\frac{1}{2\pi} \int_0^{2\pi} \frac{e^{i\theta'} J(\rho, \tau, \theta') d\theta'}{e^{i\theta'} - k}. \tag{63}$$

Hence,

$$\Psi = \left\{ \frac{1}{2\pi} \int_0^{2\pi} e^{i\theta} J(\rho, \tau, \theta) d\theta \right\} \frac{1}{k} + O\left(\frac{1}{k^2}\right), \quad k \to \infty. \tag{64}$$

Substituting this expression in Eq. (27), we find that the $O(1)$ term of Eq. (27) yields (21).

4 SRT for PET

According to Eq. (9), the inverse Radon transform can be written in the form

$$f(x_1, x_2) = \frac{1}{2i\pi} \left(\frac{\partial}{\partial x_1} - i \frac{\partial}{\partial x_2} \right) \int_0^{2\pi} e^{i\theta} F(\rho, \theta) \Big|_{\rho = x_2 \cos\theta - x_1 \sin\theta} d\theta, \tag{65}$$

where $F(\rho,\theta)$ denotes half the Hilbert transform of $\hat{f}(\rho,\theta)$ with respect to ρ, namely,

$$F(\rho,\theta) \equiv \frac{1}{2\pi} \oint_{-\infty}^{\infty} \frac{\hat{f}(r,\theta)}{r-\rho} dr, \qquad -\infty < \rho < \infty, \qquad 0 \le \theta < 2\pi. \tag{66}$$

Inserting the operator $\left(\frac{\partial}{\partial x_1} - i\frac{\partial}{\partial x_2}\right)$ inside the integral in the right- hand side of Eq. (65), we find

$$\left(\frac{\partial}{\partial x_1} - i\frac{\partial}{\partial x_2}\right) F(x_2\cos\theta - x_1\sin\theta,\theta) = -(\sin\theta + i\cos\theta)\left.\frac{\partial F(\rho,\theta)}{\partial\rho}\right|_{\rho = x_2\cos\theta - x_1\sin\theta}. \tag{67}$$

Using Eq. (67) in Eq. (65), we find

$$f(x_1,x_2) = -\frac{1}{2\pi}\int_0^{2\pi}\left[\frac{\partial F(\rho,\theta)}{\partial\rho}\right]_{\rho = x_2\cos\theta - x_1\sin\theta} d\theta, \qquad -\infty < x_1, x_2 < \infty. \tag{68}$$

For the numerical calculation of $F(\rho,\theta)$, we assume that $\hat{f}(\rho,\theta)$ has support in the interval $-1 \le \rho \le 1$ and that $\hat{f}(\rho,\theta)$ is given for every θ at the n points $\{\rho_i\}_1^n$. We denote the value of $\hat{f}$ at ρ_i by $\widehat{f_i}$, i.e.,

$$\hat{f_i} = \hat{f}(\rho_i,\theta), \qquad \rho_i \in [-1,1], \qquad 0 \le \theta < 2\pi, \quad i = 1,\ldots,n. \tag{69}$$

Furthermore, we assume that

$$\hat{f}(-1,\theta) = \hat{f}(1,\theta) = 0. \tag{70}$$

In the interval $\rho_i \le \rho \le \rho_{i+1}$, we approximate $\hat{f}(\rho,\theta)$ by cubic splines:

$$\hat{f}(\rho,\theta) = a_i(\theta) + b_i(\theta)\rho + c_i(\theta)\rho^2 + d_i(\theta)\rho^3, \quad \rho_i \le \rho \le \rho_{i+1}, \quad 0 \le \theta < 2\pi, \quad i = 1,\ldots,n, \tag{71}$$

with $\{a_i(\theta), b_i(\theta), c_i(\theta), d_i(\theta)\}_1^n$ given by the following expressions:

$$a_i(\theta) = \frac{\rho_{i+1}\hat{f}_i - \rho_i\hat{f}_{i+1}}{\Delta_i} + \frac{\hat{f}_i''}{6}\left(-\rho_{i+1}\Delta_i + \frac{\rho_{i+1}^3}{\Delta_i}\right) + \frac{\hat{f}_{i+1}''}{6}\left(\rho_i\Delta_i - \frac{\rho_i^3}{\Delta_i}\right), \tag{72}$$

$$b_i(\theta) = \frac{\hat{f}_{i+1} - \widehat{f_i}}{\Delta_i} - \frac{\hat{f}_i''}{6}\left(-\Delta_i + \frac{3\rho_{i+1}^2}{\Delta_i}\right) + \frac{\hat{f}_{i+1}''}{6}\left(-\Delta_i + \frac{3\rho_i^2}{\Delta_i}\right), \tag{73}$$

$$c_i(\theta) = \frac{1}{2\Delta_i}\left(\rho_{i+1}\hat{f}_i'' - \rho_i\,\hat{f}_{i+1}''\right), \tag{74}$$

$$d_i(\theta) = \frac{\hat{f}_{i+1}'' - \hat{f}_i''}{6\Delta_i}, \tag{75}$$

where

$$\Delta_i = \rho_{i+1} - \rho_i \tag{76}$$

and $\hat{f}_i''$ denotes the second derivative of $\hat{f}(\rho,\theta)$ with respect to ρ evaluated at ρ_i, i.e.,

$$\hat{f}_i'' = \left.\frac{\partial^2 \hat{f}(\rho,\theta)}{\partial\rho^2}\right|_{\rho=\rho_i}, \quad i = 1,\ldots,n. \tag{77}$$

It is straightforward to establish the identity

$$\frac{\partial}{\partial\rho}\oint_{\rho_i}^{\rho_{i+1}} \frac{h(r)}{r-\rho}dr = \frac{h(\rho_{i+1})}{\rho-\rho_{i+1}} - \frac{h(\rho_i)}{\rho-\rho_i} + \oint_{\rho_i}^{\rho_{i+1}} \frac{\frac{\partial h(r)}{\partial r}}{r-\rho}dr, \quad -1 \le \rho \le 1. \tag{78}$$

Employing this identity for the function $\hat{f}(\rho,\theta)$ and recalling Eqs. (69) and (70), we find

$$\frac{\partial}{\partial\rho}\oint_{-1}^{1} \frac{\hat{f}(r,\theta)}{r-\rho}dr = \sum_{i=1}^{n-1}\oint_{\rho_i}^{\rho_{i+1}} \frac{\frac{\partial \hat{f}(r,\theta)}{\partial r}}{r-\rho}dr. \tag{79}$$

Equation (71) implies

$$\oint_{\rho_i}^{\rho_{i+1}} \frac{\frac{\partial \hat{f}(r,\theta)}{\partial r}}{r-\rho}dr = b_i(\theta)\oint_{\rho_i}^{\rho_{i+1}} \frac{dr}{r-\rho} + 2c_i(\theta)\oint_{\rho_i}^{\rho_{i+1}} \frac{r\,dr}{r-\rho} + 3d_i(\theta)\oint_{\rho_i}^{\rho_{i+1}} \frac{r^2 dr}{r-\rho}. \tag{80}$$

The integrals appearing in this equation can be evaluated by employing the following identities:

$$\oint_{\rho_i}^{\rho_{i+1}} \frac{dr}{r-\rho} = I_i(\rho), \tag{81}$$

$$\oint_{\rho_i}^{\rho_{i+1}} \frac{rdr}{r-\rho} = \Delta_i + \rho I_i(\rho), \tag{82}$$

$$\oint_{\rho_i}^{\rho_{i+1}} \frac{r^2 dr}{r-\rho} = \frac{1}{2}(\rho_{i+1}^2 - \rho_i^2) + \rho\Delta_i + \rho^2 I_i(\rho), \tag{83}$$

where

$$I_i(\rho) = \ln\left|\frac{\rho_{i+1}-\rho}{\rho_i-\rho}\right|. \tag{84}$$

Using Eqs. (80)–(84), Eq. (79) becomes

$$\frac{\partial}{\partial\rho}\oint_{-1}^{1}\frac{\hat{f}(r,\theta)}{r-\rho}dr = C(\theta) + 3\left(\sum_{i=1}^{n-1} d_i(\theta)\Delta_i\right)\rho + \sum_{i=1}^{n-1} D_i(\rho)\ln\left|\frac{\rho_{i+1}-\rho}{\rho_i-\rho}\right|, \tag{85}$$

where $C(\theta)$ and $\{D_i(\rho,\theta)\}_1^{n-1}$ are defined by the equations

$$C(\theta) = \sum_{i=1}^{n-1}\left[2c_i(\theta)\Delta_i + \frac{3}{2}d_i(\theta)\left(\rho_{i+1}^2 - \rho_i^2\right)\right], \tag{86}$$

$$D_i(\rho,\theta) = b_i(\theta) + 2c_i(\theta)\rho + 3d_i(\theta)\rho^2, \quad \rho_i \le \rho \le \rho_{i+1}, \quad i = 1,\ldots,n. \tag{87}$$

After simplifying Eq. (85), we find

$$\begin{aligned}\frac{\partial F(\rho,\theta)}{\partial\rho} = \frac{1}{2\pi}\Bigg\{&C(\theta) + \frac{1}{2}\left(\hat{f}_n'' - \hat{f}_1''\right)\rho + D_{n-1}(\rho,\theta)\ln|\rho-\rho_n| - D_1(\rho,\theta)\\ &\ln|\rho-\rho_1| + \sum_{i=1}^{n-2}[D_i(\rho,\theta) - D_{i+1}(\rho,\theta)]\ln|\rho-\rho_{i+1}|\Bigg\},\\ &1 \le \rho \le 1,\ 0 \le \theta < 2\pi.\end{aligned} \tag{88}$$

In order to eliminate the logarithmic singularities at $\rho = \{\rho_1\}_1^n$, we impose the following n equations:

$$D_i(\rho_{i+1},\theta) = D_{i+1}(\rho_{i+1},\theta), \quad i = 1,\ldots,n-2, \quad 0 \le \theta < 2\pi, \tag{89}$$

$$D_1(\rho_1,\theta) = D_{n-1}(\rho_n,\theta) = 0. \tag{90}$$

In summary, the *inverse Radon transform* of a function $\hat{f}(\rho,\theta)$ approximated by the cubic spline expression (71) can be written in the form

$$f(x_1,x_2) = -\frac{1}{4\pi^2}\int_0^{2\pi}\left\{C(\theta)+\frac{1}{2}\left(\hat{f}_n''-\hat{f}_1''\right)\rho + D_{n-1}(\rho,\theta)\right.$$
$$\ln|\rho-\rho_n| - D_1(\rho,\theta)\ln|\rho-\rho_1|$$
$$\left.+\sum_{i=1}^{n-2}[D_i(\rho,\theta)-D_{i+1}(\rho,\theta)]\ln|\rho-\rho_{i+1}|\right\}d\theta, \tag{91}$$

where $C(\theta)$ and $\{D_i\}_1^{n-1}$ are defined by Eqs. (86) and (87) in terms of $\{b_i(\theta),c_i(\theta),d_i(\theta)\}_1^n$, which are defined by Eqs. (73)–(75) via $\left\{\hat{f}_i \text{ and } \hat{f}_i''\right\}_1^n$. It is assumed that the functions $\left\{\hat{f}_i(\theta)\right\}_1^n$ are given, whereas $\left\{\hat{f}_i''(\theta)\right\}_1^n$ can be computed in terms of $\hat{f}_i(\theta)$ by solving Eqs. (89) and (90).

In the construction of the so-called "natural" splines, one requires continuity of the first derivative, as well as the conditions $\hat{f}_1''=\hat{f}_n''=0$. The former requirement implies Eq. (89), which eliminates the logarithmic singularities at the interior points $\rho=\rho_i, \quad i=2,\ldots,n-1$. In order to eliminate the logarithmic singularities at the end points $\rho_1=-1$ and $\rho_n=1$, we impose Eq. (90) (instead of $\hat{f}_1''=\hat{f}_n''=0$). In this way, we construct a set of splines "custom-made" for the evaluation of the Hilbert transform.

For a discrete number N of projection angles θ, Eq. (68) yields

$$f(x_1,x_2) \sim -\frac{1}{4\pi^2 N}\left\{\sum_{j=0}^{N-1} G\left(x_1,x_2,\frac{2\pi j}{N}\right)+\frac{1}{2}G(x_1,x_2,0)+\frac{1}{2}G(x_1,x_2,2\pi)\right\}, \tag{92}$$

where $G(x_1,x_2,\theta)$ denotes the evaluation of the RHS of Eq. (88) at $\rho=x_2\cos\theta - x_1\sin\theta$.

Comparison between FBP and SRT for PET

FBP is reviewed in several publications; see, for example, [21–27]. Here, we only note that the inverse Radon transform implemented via the FBP algorithm can be expressed by the following formula [28]:

$$f(x_1,x_2) = \frac{1}{2\pi N}\sum_{n=0}^{N-1}\mathcal{F}^{-1}\left[S(\xi_\rho,\theta_n)\times H(\xi_\rho)\right], \tag{93}$$

where $\mathcal{F}^{-1}$ denotes the inverse Fourier transform in the variable ξ and $S(\xi_\rho,\theta)$ is the sinogram in the spatial frequency domain given by the expression

$$S(\xi_\rho,\theta) = \mathcal{F}\{\hat{f}(\rho,\theta)\}, \tag{94}$$

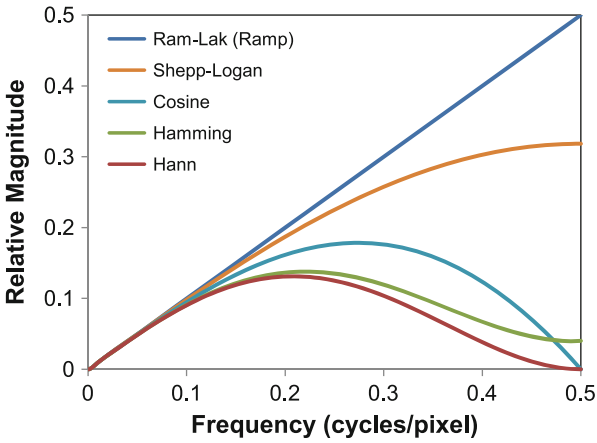


Fig. 3 The product of the ramp function with typical filters used with FBP. The cutoff frequency ξ_c is set at 0.5 cycles/pixel

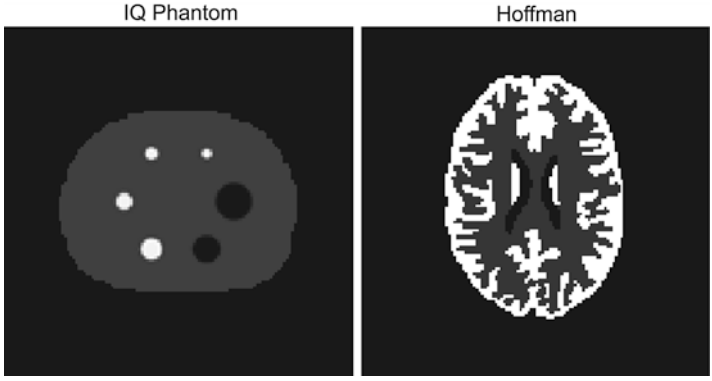


Fig. 4 The two phantoms used for the simulation studies

with $\mathcal{F}$ denoting the direct Fourier transform. The function $H(\xi_\rho)$ appearing in Eq. (93) denotes the product of the ramp function $|\xi|$ by some appropriate filter function. Figure 3 depicts $H(\xi_\rho)$ for some commonly used filters in the FBP reconstruction algorithm.

In what follows, we compare FBP (with ramp filter) with SRT using certain well-known "quality measures." These rigorous comparisons appear to indicate that SRT has certain advantages. Further comparisons can be found in [9] and [7].

Simulated Data

We will employ two well-known phantoms, namely, an image quality (IQ) phantom and a slice of a digital 3D Hoffman phantom [29]. These two phantoms are displayed in Fig. 4.

The IQ phantom, which simulates the human torso, will be used in order to establish how well each algorithm can determine hot and cold lesions of variable

size inside a warm background. It consists of two circular cold regions (with diameters of 38 and 32 mm) and four circular hot regions (with diameters of 25, 19, 15, and 12 mm) inside a larger warm region that simulates the background. The radioactive concentration ratio (RCR) between hot regions and the surrounding warm background is 4:1 for the three hot regions.

The Hoffman phantom simulates a cerebral PET study. It contains a complicated radioactivity distribution within small anatomical features; thus, it allows the investigation of the performance of each algorithm in a more realistic situation. The Hoffman phantom contains three distinct radioactive regions: gray matter (GM), white matter (WM), and cerebrovascular fluid (CSF). The RCR between GM and WM is 5:1. The radioactivity concentration in the CSF region is zero.

The GE Discovery ST PET scanner is simulated by employing STIR. Details of the scanner can be found elsewhere [30]. The two phantoms are placed in the center of the imaging system, and 2D projections are generated in STIR using a ray-tracing technique. Scatter and attenuation are not modeled. The relevant sinograms provide the noiseless PET measurements. For each noiseless sinogram, 20 Poisson noise realizations are generated at 5 different levels (NL1–NL5), where NL5 corresponds to the highest noise level applied.

The SRT algorithm constructs an image in a raster scan format, by scanning all pixel locations (x_1, x_2) and then calculating the integral over θ of the derivative of the half-Hilbert transform, which is approximated by Eq. (88). In the important case that the boundary of the object being imaged is convex, a pixel which is outside the object and hence has zero value, can be singled out from the sinogram by first identifying the detector locations for all angles θ that receive contribution from this pixel; then, for every (x_1, x_2), if there is even one θ such that $\hat{f}(\rho, \theta) = 0$, it follows that $f(x_1, x_2)$ must be zero. Using this condition, we can restrict the reconstruction process only to pixels within the object boundary and hence obtain a "cleaner" reconstructed image.

By restricting reconstruction within the object boundaries (described above) and by employing certain symmetry conditions (see [7]), it is possible to decrease substantially the reconstruction time. The actual time depends on the size of the sinogram, the reconstruction grid, and the size and the complexity of the object being imaged. For example, for the simulated Hoffman phantom (sinograms of size 221×210 and 221×221 reconstruction grid), the reconstruction time is about 2.1 s per sinogram, executed on a PC with Intel® Core™i7-920 Processor. We note that no parallel programming or other accelerating techniques are employed, which of course will decrease further the reconstruction time. The corresponding reconstruction time for FBP in STIR is about 0.3 s.

In order to determine the quality of the reconstructed images, the contrast and bias for the IQ and Hoffman phantom are calculated. The contrast for the hot and cold regions are calculated according to [30]. The bias is defined as the mean deviation over all realizations of the mean pixel value within a region of interest (ROI) from the actual activity concentration, i.e.,

$$\text{bias} = \left\{ \frac{1}{R} \sum_{r=1}^{R} \overline{X}_r \right\} - X_{\text{true}}, \tag{95}$$

where R is the total number of realizations, X_{true} is the true activity concentration, and $\overline{X}_r$ is the mean activity concentration within an ROI of realization r with M number of pixels; $\overline{X}_r$ is given by

$$\overline{X}_r = \frac{1}{N} \sum_{i=1}^{N} X_i . \tag{96}$$

The coefficient of variations (COV) is calculated using the expression

$$\text{COV} = \frac{\sigma}{m}, \tag{97}$$

where σ and m are the standard deviation and the mean of the measured activity in the background ROI, averaged over all realizations.

For the IQ simulated phantom, comparisons between SRT and FBP reconstructed images with no noise as well as with noise (NL2, NL4, and NL5) are shown in Fig. 5. For economy of presentation, images from NL1 and NL3 are not shown. The noisy images presented are representative reconstructions of one realization at the specific noise level. Both SRT and FBP can generate negative values in pixels (for cases where the value of the original phantom is very low or zero). In all images presented, the all-black color corresponds to zero values, whereas the white color represents the maximum value of the distribution.

Although the reconstructions from both methods appear similar, there exist two main differences: From visual inspection, it is clear that there exist differences in the noise texture between the SRT and the FBP reconstructions. Specifically, the reconstructions obtained from SRT appear more noisy than those obtained from FBP at every noise level. Furthermore, the SRT reconstructions are completely clear from streak artifacts outside the object, whereas some small streak artifacts are present in the FBP reconstructions.

The contrast, C_{hot}, for the two smallest hot spheres of the IQ phantom (15 and 12 mm) as a function of COV is presented in Fig. 6. The SRT algorithm exhibits higher contrast in all three lesions independently of noise level. This advantage in contrast increases as the size of the lesion decreases. Indeed, we observe no differences in C_{cold} for the 38-mm and 32-mm cold spheres but a small improvement for the SRT images in C_{hot} for the 25-mm lesion and larger improvement for the 19-mm lesion.

The percent bias generated by the reconstruction algorithms for the two smaller hot lesions as a function of noise level is presented in Fig. 6. The bias is negative in all cases. There are no considerable differences in bias between SRT and FBP for the cold lesion and for the largest hot lesion. The bias had small variations as

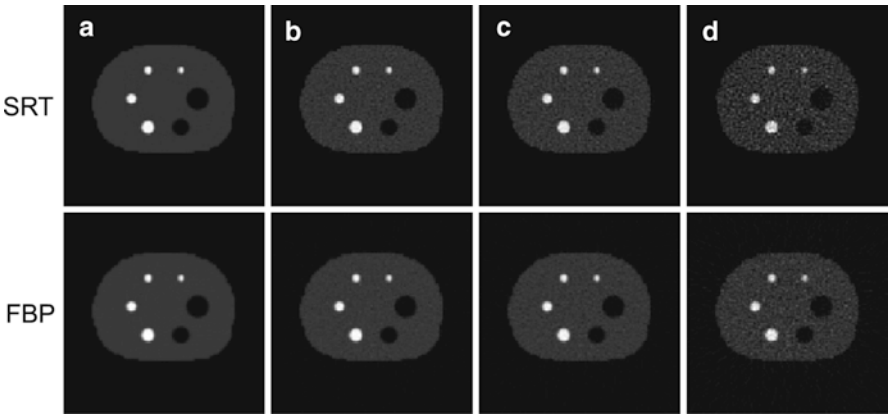


Fig. 5 Reconstructed images of the simulated IQ phantom at various noise levels (the noise level increases moving from left to right): (**a**) no noise, (**b**) noise level 2, (**c**) noise level 4, and (**d**) noise level 5

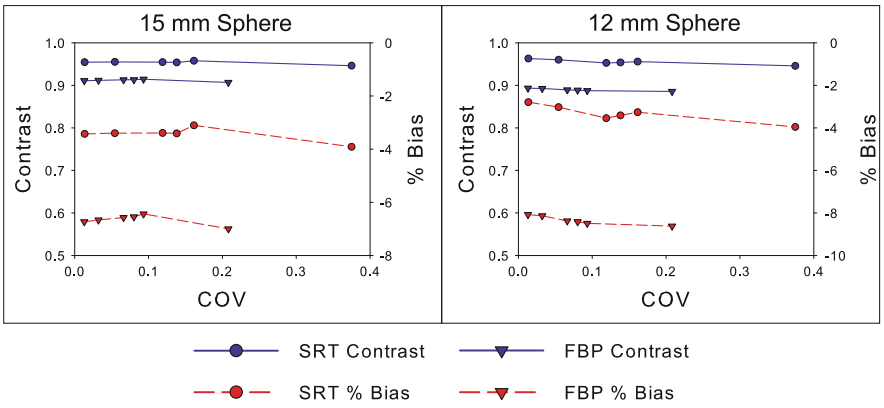


Fig. 6 Contrast and bias vs. COV for the two reconstruction algorithms obtained from the reconstructed images of the simulated IQ Phantom. Note that the leftmost data point in each line curve corresponds to the noiseless case, while the rightmost data point corresponds to the NL5 case

a function of COV for both SRT and FBP. FBP appears to give a higher bias in all cases reaching about 9 % for the 12-mm lesion. The percent bias for the FBP images increases as the lesion size decreases. The percent bias for SRT has small variations as a function of lesion size. The situation is similar for the 19-mm hot lesion.

For the Hoffman simulated phantom, comparisons between SRT and FBP reconstructed images with no noise as well as with noise (NL2, NL4, and NL5) are shown in Fig. 7. The noisy images presented are representative reconstructions of one realization at the specific noise level. All anatomical features of this phantom can clearly be identified by both algorithms for all selected noise levels. Small streak artifacts outside the object are present in the FBP reconstructions, whereas the SRT reconstruction provides images with no such artifacts.

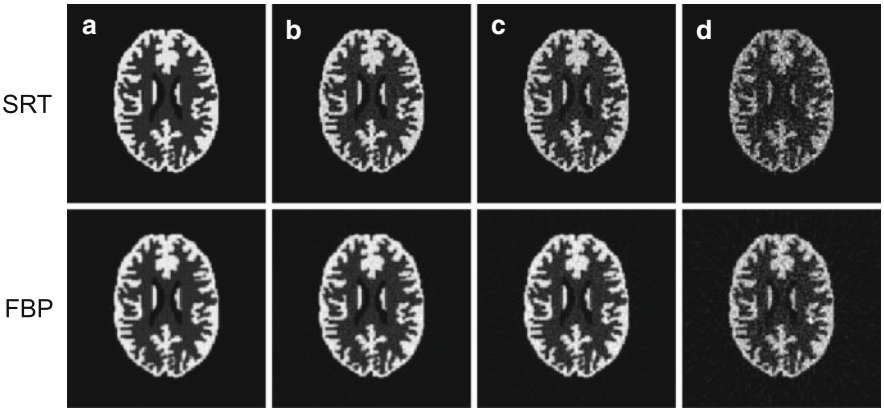


Fig. 7 Reconstructed images of the simulated Hoffman phantom at various noise levels: (**a**) no noise, (**b**) noise level 2, (**c**) noise level 4, and (**d**) noise level 5

Contrast plots between GM/CSF and WM/CSF as a function of noise level are presented in Fig. 8a, b. There exists a small improvement in contrast for the SRT algorithm, particularly for the case of the WM. RCR plots comparing the two reconstruction algorithms as a function of noise level are presented in Fig. 8c. The RCR calculations between GM and WM suggest that FBP slightly underestimates the actual RCR value (dotted line), whereas the RCR calculated from SRT reconstructions is closer to the actual value. The percent bias for the GM as a function of noise level is depicted in Fig. 8d. There is a negative bias in both algorithms, similar to the case of the IQ phantom. The bias is 4 % for the FBP and about 2 % for the SRT algorithm. The bias for the GM is under 0.8 % for both algorithms.

Real Data

Real-data acquisitions are performed using a commercial ARGUS-CT small-animal PET-CT system (SEDECAL S.A., Madrid, Spain). The PET tomograph of this system is identical to the GE Healthcare eXplore VISTA small-animal PET scanner which is described elsewhere [31].

The following two phantoms are used: (a) a NEMA image quality phantom and (b) an in-house Derenzo phantom. A NEMA phantom, designed in accordance to the specifications of the NEMA NU 4-2008 quality phantom [32], is employed in order to determine the noise and contrast properties of each algorithm. This phantom is separated into three main parts: a fillable cylindrical region 30 mm in diameter and 30 mm in length; a solid region with 5 fillable rods with dimensions of 1, 2, 3, 4, and 5 mm each; and a uniform region with two cold region chambers 8 mm in diameter. Schematics of the three distinct parts of the NEMA phantom are shown in Fig. 9. The entire phantom is filled with 15.8 MBq of ^{18}F aqueous solution, and one of the

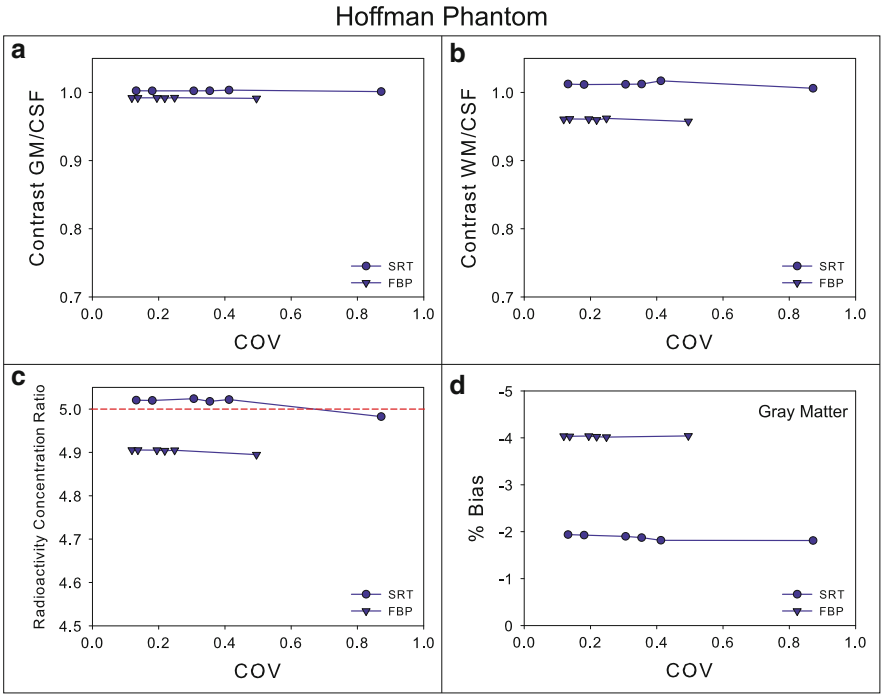


Fig. 8 Contrast, RCR, and bias vs. COV comparisons between the two reconstruction algorithms for the simulated Hoffman phantom. Contrast was determined for GM with respect to CSF (**a**) and for WM with respect to CSF (**b**). RCR between GM and WM was also determined (**c**). The *dashed line* indicates the actual RCR between GM and WM of the simulated Hoffman phantom being imaged. The bias is presented as a percentage of the true activity concentration of the lesion being measured (**d**). Note that the leftmost data point in each line curve corresponds to the noiseless case, while the rightmost data point corresponds to the NL5 case

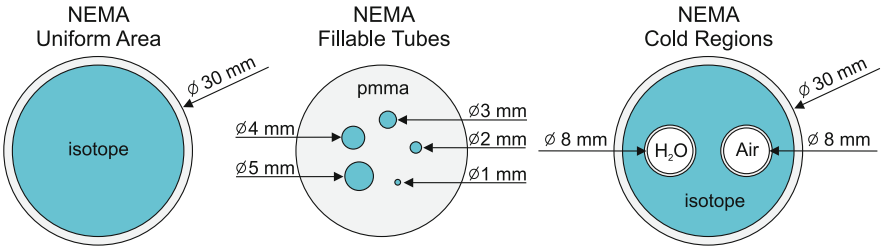


Fig. 9 Schematics of the three distinct parts of the NEMA NU 4-2008 phantom used for the real-data studies

8-mm cold chambers is filled with nonradioactive water while the other one remains with air. A 30-min PET scan is acquired in two bed positions.

An in-house Derenzo phantom is used in order to test the resolution limitations of each algorithm. The Derenzo phantom consists of 31 microcapillaries arranged

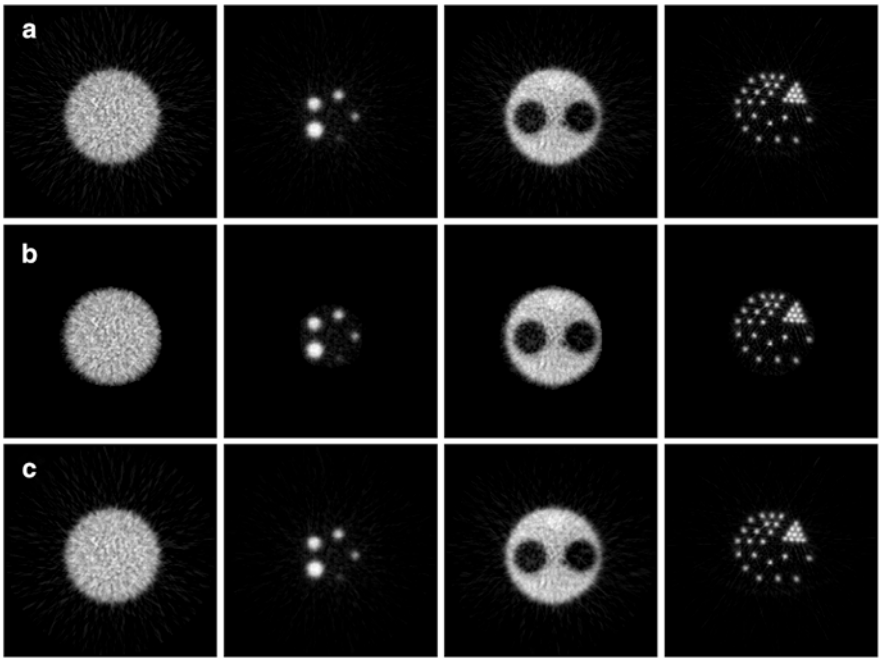


Fig. 10 Reconstructions of three slices of the NEMA NU 4-2008 image quality phantom and a slice of the Derenzo phantom, acquired by the ARGUS-CT small-animal PET-CT system: (**a**) SRT with no thresholding, (**b**) SRT with thresholding, and (**c**) FBP with a ramp filter. Note that the 8-mm cold chamber on the left of the NEMA 2 slice is filled with nonradioactive water and the one on the right with air

in six different sectors. The capillaries are separated by 2, 3, 4, 5, 6, and 8 mm, respectively, and there is no material between them. The phantom is filled with 5.6 MBq of ^{18}F aqueous solution, and a 60-min PET scan is performed.

Reconstructed images obtained via SRT and FBP (including SRT without sinogram thresholding) of three slices of the NEMA NU 4-2008 image quality phantom are presented in Fig. 10. Small streak artifacts are present in both SRT and FBP reconstructions. These artifacts are reduced in SRT after applying sinogram thresholding.

COV and contrast calculations using the uniform slice and the cold chamber slice of the NEMA phantom are presented in Fig. 11a, b. The SRT reconstructed images exhibit slightly higher COV values in comparison to FBP. The contrast values for both the water and air chambers are similar for both algorithms.

Reconstructed images of the Derenzo phantom are shown in Fig. 10. All reconstructed circular sources are clearly visible with both methods. Fig. 11c illustrates the contrast for the various sectors of the Derenzo phantom calculated from the SRT and FBP reconstructed images. The contrast in the SRT reconstructed images is higher than the contrast of the FBP images; the difference in contrast between SRT

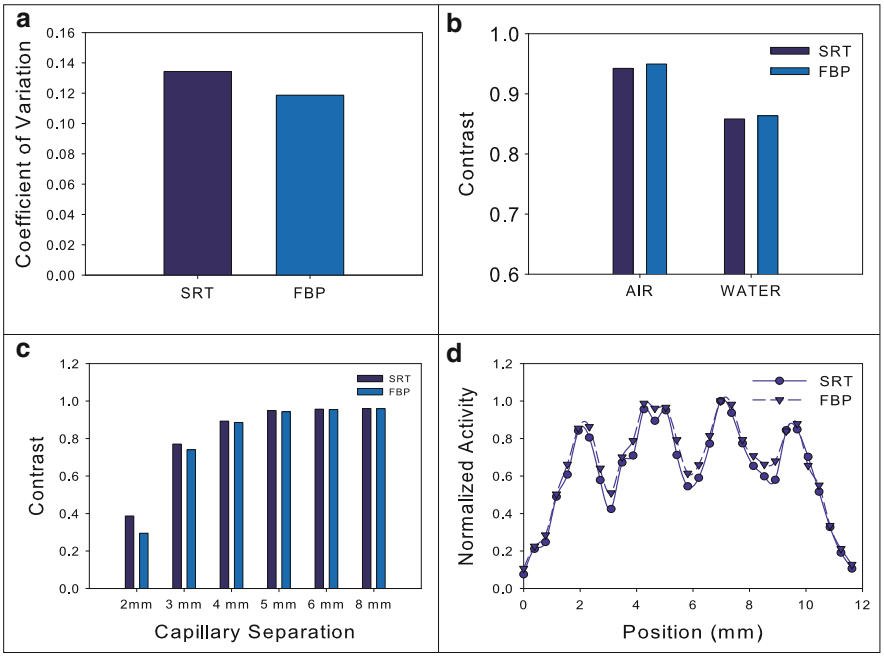


Fig. 11 Results from analyzing real NEMA NU 4-2008 and Derenzo phantom images: (**a**) COV, (**b**) contrast for water to background and air to background, (**c**) contrast for the different sections of the Derenzo phantom, and (**d**) line profiles obtained along the 2-mm separated capillaries of the Derenzo phantom. The *dashed lines* correspond to the normalized profile values through the actual phantom

and FBP images becomes larger as the center-to-center spacing between the lesions of the Derenzo phantom becomes smaller. Figure 11d illustrates the line profiles obtained along the 2-mm separated capillaries of the phantom. SRT resolves the 2-mm separated holes slightly better than FBP.

5 SRT for SPECT

In order to derive the final formula for IART, we first derive the following simplification of Eq. (21):

$$f(x_1,x_2)=\frac{1}{2\pi}\int_0^{2\pi}\exp\{M(x_1,x_2,\theta)\}\left\{\left[-\cos\theta\frac{\partial M(x_1,x_2,\theta)}{\partial x_2}+\sin\theta\frac{\partial M(x_1,x_2,\theta)}{\partial x_1}\right]G(\rho,\theta)-\frac{\partial G(\rho,\theta)}{\partial\rho}\right\}_{\rho=x_2\cos\theta-x_1\sin\theta}d\theta, \quad (98)$$

where the functions M and G are defined by

$$M(x_1, x_2, \theta) = \int_{\tau}^{\infty} \mu(s\cos\theta - \rho\sin\theta, s\sin\theta + \rho\cos\theta)ds \Big|_{\substack{\tau = x_2\sin\theta + x_1\cos\theta \\ \rho = x_2\cos\theta - x_1\sin\theta}} \tag{99}$$

and

$$G(\rho,\theta) = \exp\left\{-\frac{\hat{\mu}(\rho,\theta)}{2}\right\}\left[\cos(F(\rho,\theta))G^C(\rho,\theta) + \sin(F(\rho,\theta))G^S(\rho,\theta)\right], \tag{100}$$

with F, G^C, G^S defined by

$$F(\rho,\theta) = \frac{1}{2\pi}\oint_{-\infty}^{\infty}\frac{\hat{\mu}(r,\theta)}{r-\rho}dr, \tag{101}$$

$$G^C(\rho,\theta) = \frac{1}{2\pi}\oint_{-\infty}^{\infty}\exp\left\{\frac{\hat{\mu}(r,\theta)}{2}\right\}\cos(F(r,\theta))\hat{f}_\mu(r,\theta)\frac{dr}{r-\rho}, \tag{102}$$

$$G^S(\rho,\theta) = \frac{1}{2\pi}\oint_{-\infty}^{\infty}\exp\left\{\frac{\hat{\mu}(r,\theta)}{2}\right\}\sin(F(r,\theta))\hat{f}_\mu(r,\theta)\frac{dr}{r-\rho}, \\ -\infty < \rho < \infty, \qquad 0 \le \theta < 2\pi. \tag{103}$$

In order to derive Eq. (98), we insert the operator $\partial_{x_1} - i\,\partial_{x_2}$ inside the integral in the RHS of (21) and use the identity

$$\begin{aligned}(\partial_{x_1} - i\,\partial_{x_2}) &= \frac{\partial x_1}{\partial\tau}\partial\tau + \frac{\partial x_1}{\partial\rho}\partial\rho - i\left(\frac{\partial x_2}{\partial\tau}\partial\tau + \frac{\partial x_2}{\partial\rho}\partial\rho\right) \\ &= (\cos\theta - i\,\sin\theta)\,\partial\tau + (-\sin\theta - i\,\cos\theta)\,\partial\rho = e^{-i\theta}(\partial_\tau - i\,\partial_\rho).\end{aligned} \tag{104}$$

Then, Eq. (21) becomes

$$(\partial_{x_1} - i\,\partial_{x_2})J(x_1,x_2,\theta) = e^{-i\theta}e^{M}\left\{\left[-\mu(x_1,x_2)L_\mu\hat{f}_\mu + \left(\sin\theta\frac{\partial M}{\partial x_1} - \cos\theta\frac{\partial M}{\partial x_2}\right)\right]\left(iL_\mu\hat{f}_\mu\right) - \frac{\partial}{\partial\rho}\left(iL_\mu\hat{f}_\mu\right)\right\}_{\substack{\tau = x_2\sin\theta + x_1\cos\theta \\ \rho = x_2\cos\theta - x_1\sin\theta}}. \tag{105}$$

The contribution to Eq. (21) of the first term of the RHS of Eq. (105) vanishes. Indeed, this contribution equals

$$-\mu(x_1,x_2)\int_0^{2\pi}\exp\{M(x_1,x_2,\theta)\}\,(L_\mu\hat{f}_\mu)(\rho,\theta)\Big|_{\rho=x_2\cos\theta-x_1\sin\theta}d\theta=$$
$$-\mu(x_1,x_2)\int_0^{2\pi}J(x_1,x_2,\theta)d\theta. \qquad (106)$$

But Eqs. (1.8) and (2.9) of [33] evaluated at $\lambda=0$ imply

$$\int_0^{2\pi}J(x_1,x_2,\theta)d\theta=0. \qquad (107)$$

Thus, using Eq. (105) in Eq. (21), we find Eq. (98) provided that the following identity is valid:

$$i\left(L_\mu\hat{f}_\mu\right)(\rho,\theta)=2G(\rho,\theta). \qquad (108)$$

In order to derive this equation, we will use the equations

$$\exp\{P^-\hat{\mu}\}=\exp\left\{-\frac{\hat{\mu}}{2}-iF\right\},\qquad \exp\{P^+\hat{\mu}\}=\exp\left\{\frac{\hat{\mu}}{2}-iF\right\},$$

where F denotes half the Hilbert transform of $\hat{\mu}$ in the variable ρ; see Eq. (101). Hence,

$$\exp\{P^-\hat{\mu}\}\,P^-\exp\{-P^-\hat{\mu}\}\,\hat{f}_\mu=$$
$$\exp\left\{-\frac{\hat{\mu}}{2}-iF\right\}\left[-\frac{1}{2}\exp\left\{\frac{\hat{\mu}}{2}+iF\right\}\hat{f}_\mu+\frac{1}{2i}H\left(\exp\left\{\frac{\hat{\mu}}{2}+iF\right\}\hat{f}_\mu\right)\right] \qquad (109)$$

and

$$\exp\{-P^+\hat{\mu}\}\,P^+\exp\{P^+\hat{\mu}\}\,f_\mu=$$
$$\exp\left\{-\frac{\hat{\mu}}{2}+iF\right\}\left[\frac{1}{2}\exp\left\{\frac{\hat{\mu}}{2}-iF\right\}\hat{f}_\mu+\frac{1}{2i}H\left(\exp\left\{\frac{\hat{\mu}}{2}-iF\right\}\hat{f}_\mu\right)\right], \qquad (110)$$

where H denotes the Hilbert transform in the variable ρ defined in Eq. (11). Using the above equations in Eq. (23) and simplifying, we find

$$i(L_\mu\hat{f}_\mu)(\rho,\theta)=\frac{1}{2}\exp\left\{-\frac{\hat{\mu}}{2}\right\}\times$$
$$\left[\exp\{-iF\}\,H\left(\exp\left\{\frac{\hat{\mu}}{2}+iF\right\}\hat{f}_\mu\right)+\exp\{iF\}\,H\left(\exp\left\{\frac{\hat{\mu}}{2}-iF\right\}\hat{f}_\mu\right)\right] \qquad (111)$$

It is important to note that the RHS of Eq. (111) is real.

Using the definitions (102) and (103), after some simplifications, Eq. (111) becomes

$$i\left(L_\mu \hat{f}_\mu\right)(\rho,\theta) = \exp\left\{-\frac{\hat{\mu}}{2}\right\}\left[\cos(F)G^C + \sin(F)G^S\right]. \tag{112}$$

6 Conclusion

PET and SPECT constitute two important medical imaging techniques. Using these techniques, images can be obtained by employing either analytic methods or iterative methods.

Regarding analytical methods, we note that simple models for PET and SPECT can be formulated in terms of the inverse Radon transform and the IART, respectively. FBP provides the most well-known numerical implementation of the inverse Radon transform. FBP is currently used commercially for both PET and SPECT, in spite of the fact that SPECT involves the IART and not the inverse Radon transform (this is a consequence of the fact that until recently no analytic formula was available for the IART).

Here, we have reviewed a general method for deriving transform pairs and have used this method to derive the inverse Radon transform and the IART. Furthermore, we have presented a novel numerical technique based on "custom-made" cubic splines for the numerical implementation of both the above inverse transforms. Rigorous studies comparing this novel technique (called SRT) with FBP for PET suggest that SRT has certain advantages. In future clinical studies, efforts will be made to delineate this advantage in concrete clinical situations.

The main advantages of the analytical methods, like FBP and SRT, are speed and simplicity. However, in these methods, it is difficult to incorporate complex physical phenomena such as attenuation and scatter. In FBP, noise issues are treated by selecting appropriate filtering parameters, such as the roll-off and cutoff frequencies of the reconstruction filter (usually at the expense of spatial resolution). Another disadvantage of FBP is the streak artifacts that are particularly prominent near hot regions of the object. For SRT, it is possible to eliminate these effects at least outside the main region of interest and hence to obtain a "cleaner" image.

The predominant iterative algorithms are the maximum likelihood expectation maximization (MLEM) algorithm [34] and its accelerated successor the ordered subset expectation maximization (OSEM) algorithm [35]. The main advantage of the iterative algorithms is the ability to model several aspects of the imaging system, including elements of the noise characteristics, sinogram blurring due to detector crystal penetration, inter-crystal scatter, depth of interaction, and photon attenuation [36,37]. As a consequence, iterative methods can improve image quality and achieve considerable resolution recovery. However, iterative algorithms require more computing time and power, particularly when details of the physical model are

included. Iterative techniques are now in widespread use in clinical and preclinical systems. This is due to the speed improvement provided by OSEM and the recent computer hardware improvements (processing and storage).

Most commercial clinical and preclinical PET systems allow the use of either FBP or OSEM for image reconstruction. Currently, in OSEM, the main challenge is the selection of the proper number of subsets and iterations [38], as well as the choice of a suitable post-reconstruction filter (if needed). Stopping the algorithm at the proper number of iterations is important, since EM-based algorithms suffer from noise/bias trade-off. Stopping the iteration process after convergence is reached results in a noisy image, whereas stopping the process too soon results in a less noisy image which however is biased toward the image assumed at the initial step. In order to resolve this issue, several regularization schemes have been proposed [39]. In spite of these improvements, a recent dynamic brain PET study by Reilhac et al. [40] concludes that analytical methods are more robust to low-count data than iterative methods. Furthermore, the positivity constraint imposed to the sinogram and image space by the EM-based reconstruction algorithms leads to overestimation (positive bias) of the low-activity regions [40–42].

The advantage of analytical methods for low-count data will be increasingly more important. Indeed, in the 1980s, medical imaging was responsible for only about 15 % of the total radiation exposure to US population from all sources; now this proportion is 50 % [43]. In 2010, 70 million CT scans were performed in the USA, and the radiation dose from CT scans is 100–500 times those from conventional radiography. Taking into consideration that CT is now used for screening for lung cancer [44] and that the use of PET-CT and SPECT-CT is expanding, it is imperative to be able to produce images with the least possible radiation.

7 Cross-References

▸ Tomography
▸ EM Algorithms

References

1. U-King-Im, J.M., Young, V., Gillard, J.H.: Carotid-artery imaging in the diagnosis and management of patients at risk of stroke. Lancet Neurol. **8**(6), 569–580 (2009)
2. Hutchins, G.D., Miller, M.A., Soon, V.C., Receveur, T.: Small animal PET imaging. ILAR J. **49**(1), 54–65 (2008)
3. Cherry, S.R., Gambhir, S.S.: Use of positron emission tomography in animal research. ILAR J. **42**(3), 219–232 (2001)
4. Lucas, A.J., Hawkes, R.C., Ansorge, R.E., Williams, G.B., Nutt, R.E., Clark, J.C., Fryer, T.D., Carpenter, T.A.: Development of a combined microPET-MR system. Technol. Cancer Res. Treat. **5**(4), 337–341 (2006)

5. Cherry, S.R.: Fundamentals of positron emission tomography and applications in preclinical drug development. J. Clin. Pharmacol. **41**(5), 482–491 (2001)
6. Zaidi, H., Hasegawa, B.H.: The problem of photon attenuation in emission tomography. In: Zaidi, H. (ed.) Quantitative Analysis in Nuclear Medicine Imaging, p. 167. Springer, New York (2006)
7. Kastis, G.A., Kyriakopoulou, D., Gaitanis, A., Fernandez, Y., Hutton, B.F., Fokas, A.S.: Evaluation of the spline reconstruction technique for PET. Med. Phys. **41**, 042501 (2014). doi:http://dx.doi.org/10.1118/1.4867862
8. Thielemans, K., Tsoumpas, C., Mustafovic, S., Beisel, T., Aguiar, P., Dikaios, N., Jacobson, M.W.: STIR: software for tomographic image reconstruction Release 2. Phys. Med. Biol. **57**(4), 867–883 (2012)
9. Kastis, G.A., Gaitanis, A., Skouras, T., Fokas, A.S.: Evaluation of a spline reconstruction technique for SPECT: comparison with FBP and OSEM. In: Conference Record of the 2011 IEEE Nuclear Science Symposium and Medical Imaging Conference, Oct 2011
10. Fokas, A.S., Gelfand, I.M.: Integrability of linear and nonlinear evolution equations, and the associated nonlinear Fourier transforms. Lett. Math. Phys. **32**(3), 189–210 (1994)
11. Pauwels, E.K., Ribeiro, M.J., Stoot, J.H., McCready, V.R., Bourguignon, M., Mazière, B.: FDG accumulation and tumor biology. Nucl. Med. Biol. **25**(4), 317–322 (1998)
12. Fischer, B., Lassen, U., et al.: Preoperative staging of lung cancer with combined PET-CT. N. Engl. J. Med. **361**(1), 32–39 (2009)
13. Armitage, J.O.: Early-stage Hodgkin's lymphoma. N. Engl. J. Med. **363**(7), 653–662 (2010)
14. Baterman, R.J., Xiong, C., et al.: Clinical and biomarker changes in dominantly inherited Alzheimer's disease. N. Engl. J. Med. **367**(9), 795–804 (2012)
15. Abrams, J.: Clinical practice. Chronic stable angina. N. Engl. J. Med. **352**(24), 2524–2533 (2005)
16. Gershenwald, J.E., Ross, M.I.: Sentinel-lymph-node biopsy for cutaneous melanoma. N. Engl. J. Med. **364**(18), 1738–1745 (2011)
17. The Parkinson Study Group: Levodopa and the progression of Parkinson's disease. N. Engl. J. Med. **351**(24), 2498–2508 (2004)
18. Fokas, A.S., Novikov, R.G.: Discrete analogues of the $\overline{\partial}$ equation and of Radon transform. C. R. Acad. Sci. Paris **313**(2), 75–80 (1991)
19. Novikov, R.G.: An inversion formula for the attenuated X-ray transformation. Ark. Mat. **40**(1), 145–167 (2002)
20. Miqueles, E.X., De Pierro, A.R.: Exact analytic reconstruction in x-ray fluorescence CT and approximated versions. Phys. Med. Biol. **55**(4), 1007–1024 (2010)
21. Herman, G.T.: Fundamentals of Computerized Tomography: Image Reconstruction from Projections, 2nd edn. Springer, Dordrecht/Heidelberg/London/New York (1980)
22. Natterer, F.: The Mathematics of Computerized Tomography. Wiley, New York (1986)
23. Natterer, F., Wübbeling, F.: Mathematical Methods in Image Reconstruction. SIAM, Philadelphia (2001)
24. Parker, J.A.: Image Reconstruction in Radiology. CRC, Boca Raton (1990)
25. Epstein, C.L.: Introduction to the Mathematics of Medical Imaging. SIAM, Philadelphia (2008)
26. Kinahan, P.E., Defrise, M., Clackdoyle, R.: Analytic image reconstruction methods. In: Wernick, M.N., Aarsvold, J.N. (eds.) Emission Tomography: The Fundamentals of PET and SPECT, p. 421. Elsevier Academic, San Diego (2004)
27. Tsui, B.M.W., Frey, E.C.: Analytic image reconstruction methods in emission computed tomography. In: Zaidi, H. (ed.) Quantitative Analysis in Nuclear Medicine Imaging, p. 82. Springer, New York (2006)
28. Cherry, S.R., Dahlbom, M.: PET: physics, instrumentation, and scanners. In: Phelps, M.E. (ed.) Physics, Instrumentation, and Scanners, p. 76. Springer, New York (2010)
29. Hoffman, E.J., Cutler, P.D., Digby, W.M., Mazziotta, J.C.: 3D phantom to simulate cerebral blood flow and metabolic images for PET. IEEE Trans. Nucl. Sci. **37**(2), 616–620 (1990)

30. Bettinardi, V., Danna, M., Savi, A., Lecchi, M., Castiglioni, I., Gilardi, M.K., Bammer, H., Lucignani, G., Fazio, F.: Performance evaluation of the new whole-body PET/CT scanner: discovery ST. Eur. J. Nucl. Med. Mol. Imaging **31**(6), 867–881 (2004)
31. Wang, Y., Seidel, J., Tsui, B.M.W., Vaquero, J.J., Pomper, M.G.: Performance evaluation of the GE Healthcare eXplore VISTA dual-ring small-animal PET scanner. J. Nucl. Med. **47**(11), 1891–1900 (2006)
32. NEMA: NEMA NU 4-2008: Performance Measurements of Small Animal Positron Emission Tomographs. National Electrical Manufacturers Association, Rosslyn (2008)
33. Fokas, A.S., Iserles, A., Marinakis, V.: Reconstruction algorithm for single photon emission computed tomography and its numerical implementation. J. R. Soc. Interface **3**(6), 45–54 (2006)
34. Shepp, L.A., Vardi, Y.: Maximum likelihood reconstruction for emission tomography. IEEE Trans. Med. Imaging **1**(2), 113–121 (1982)
35. Hudson, M., Larkin, R.S.: Accelerated image reconstruction using ordered subsets of projection data. IEEE Trans. Med. Imaging **13**(4), 601–609 (1994)
36. Carson, R.E., Yan, Y., Chodkowski, B., Yap, T.K., Daube-Witherspoon, M.E.: Precision and accuracy of regional radioactivity quantitation using the maximum likelihood EM reconstruction algorithm. IEEE Trans. Med. Imag. **13**(3), 526–537 (1994)
37. Johnson, C.A., Yan, Y., Carson, R.E., Martino, R.L., Daube-Witherspoon, M.E.: A system for the 3D reconstruction of retracted-septa PET data using the EM algorithm. IEEE Trans. Nucl. Sci. **42**(4), 1223–1227 (1995)
38. De Jonge, F.A.A., Blokland, K.A.K.: Statistical tomographic reconstruction: How many more iterations to go? Eur. J. Nucl. Med. **26**(10), 1247–1250 (1999)
39. Bettinardi, V., Pagani, E., Gilardi, M., Alenius, S., Thielemans, K., Teras, M., Fazio, F.: Implementation and evaluation of a 3D one-step late reconstruction algorithm for 3D positron emission tomography brain studies using median root prior. Eur. J. Nucl. Med. Mol. Imaging **29**(1), 7–18 (2002)
40. Reilhac, A., Tomei, S., Buvat, I., Michel, C., Keheren, F., Costes, N.: Simulation-based evaluation of OSEM iterative reconstruction methods in dynamic brain PET studies. Neuroimage **39**(1), 359–368 (2008)
41. Verhaeghe, J., Reader, A.J.: AB-OSEM reconstruction for improved Patlak kinetic parameter estimation: a simulation study. Phys. Med. Biol. **55**(22), 6739–6757 (2010)
42. Bélanger, M.J., Mann, J.J., Parsey, R.V.: OS-EM and FBP reconstructions at low count rates: effect on 3D PET studies of [^{11}C] WAY-100635. Neuroimage **21**(1), 244–250 (2004)
43. Brenner, D.J.: Medical imaging in the 21st century-getting the best bang for the rad. N. Engl. J. Med. **362**(10), 943–945 (2010)
44. Kovalchik, S.A., Tammemagi, M., Berg, C.D., Caporaso, N.E., Riley, T.L., Korch, M., Silvestri, G.A., Chaturvedi, A.K., Katki, H.A.: Targeting of low-dose CT screening according to the risk of lung-cancer death. N. Engl. J. Med. **369**(3), 245–254 (2013)

Mathematics of Electron Tomography

Ozan Öktem

Contents

1 Introduction 939
2 The Transmission Electron Microscope (TEM) 940
Sample Preparation 942
3 Basic Notation and Definitions 943
4 The Forward Model 944
Illumination 945
Electron–Specimen Interaction 946
Optics 957
Detection 962
Forward Operator for Combined Phase and Amplitude Contrast 966
Forward Operator for Amplitude Contrast Only 969
Summary 970
5 Data Acquisition Geometry 971
Parallel Beam Geometries 972
Examples Relevant for ET 972
6 The Reconstruction Problem in ET 974
Mathematical Formulation 974
Notion of Solution 976
7 Specific Difficulties in Addressing the Inverse Problem 977
The Dose Problem 977
Incomplete Data, Uniqueness, and Stability 978
Nuisance Parameters 981
8 Data Pre-processing 985
Basic Pre-processing 985
Alignment 986
Deconvolving Detector Response 986
Deconvolving Optics PSF 987
Phase Retrieval 987

O. Öktem
Department of Mathematics, KTH Royal Institute of Technology, Stockholm, Sweden
e-mail: ozan@kth.se

© Springer Science+Business Media New York 2015

O. Scherzer (ed.), *Handbook of Mathematical Methods in Imaging*,
DOI 10.1007/978-1-4939-0790-8_43

9 Reconstruction Methods 988
Analytic Methods 989
Approximative Inverse 999
Iterative Methods with Early Stopping 1001
Variational Methods 1005
Other Reconstruction Schemes 1013
10 Validation 1014
11 Examples 1016
Balls 1016
Virions and Bacteriophages in Aqueous Buffer 1020
12 Conclusion 1020
Cross-References 1022
References 1022

Abstract

This survey starts with a brief description of the scientific relevance of electron tomography in life sciences followed by a survey of image formation models. In the latter, the scattering of electrons against a specimen is modeled by the Schrödinger equation, and the image formation model is completed by adding a description of the transmission electron microscope optics and detector. Electron tomography can then be phrased as an inverse scattering problem and attention is now turned to describing mathematical approaches for solving that reconstruction problem. This part starts out by explaining challenges associated with the aforementioned inverse problem, such as the extremely low signal-to-noise ratio in the data and the severe ill-posedness due to incomplete data, which naturally brings up the issue of choosing a regularization method for reconstruction. Here, the review surveys both methods that have been developed, as well as pointing to new promising approaches. Some of the regularization methods are also tested on simulated and experimental data. As a final note, this is not a traditional mathematical review in the sense that focus here is on the application to electron tomography rather than on describing mathematical techniques that underly proofs of key theorems.

Acronyms

ART	Algebraic Reconstruction Technique
CCD	Charged Coupled Device
CPMV	Cowpea Mosaic Virus
CTF	Contrast Transfer Function
EELS	Electron Energy Loss Spectroscopy
EDS	Energy-Dispersive X-ray Spectroscopy
ET	Electron (microscopy) Tomography
ELT	Electron Λ-Tomography
FBP	Filtered Back-Projection
HAADF	High-Angle Annular Dark-Field

HRTEM	High-Resolution TEM
POCS	Projection Onto Convex Sets
LDDM	Large Deformation Diffeomorphic Metric Mapping
ML	Maximum-Likelihood
ML-EM	Maximum-Likelihood Expectation Maximization
MTF	Modulation Transfer Function
PSF	Point Spread Function
SART	Simultaneous ART
SEM	Scanning Electron Microscope
SIRT	Simultaneous Iterative Reconstruction Technique
SSC	Slow-Scan CCD camera
STEM	Scanning Transmission Electron Microscope
TEM	Transmission Electron Microscope
TMV	Tobacco Mosaic Virus
TV	Total Variation
WBP	Weighted Back-Projection
WKB	Wentzel–Kramers–Brillouin

1 Introduction

Imaging is today an essential tool in many different areas of scientific and technological research. It is also widely used in investigations in fields as diverse as arts and jurisprudence. As such, it is integrated into a variety of devices and procedures routinely used in society. A prime example is microscopy that enables investigation of objects that are too small for the naked eye at a variety of scales, from atomic to sub-millimeter. Therefore, various forms of microscopy are crucial in industrial processes, in life sciences, in comparative studies of chemistry and geology, and in diagnostic medicine.

Traditional imaging techniques provide 2D images whose usage involves significant portion of interpretation since the object/phenomena under investigation almost always takes place in three spatial dimensions, plus time if temporal variation is included. 3D images are therefore to be preferred whenever possible. This applies in particular to life sciences and drug discovery, where a central topic is to understand the machinery within the cell responsible for supporting life and disease. An important part in this quest is to map the spatial and temporal arrangement of the molecules engaged in this machinery. The *structure determination problem*, which seeks to *recover the three-dimensional structure of an individual molecule at highest possible resolution in its natural environment*, has therefore come to play a central role.

Addressing the structure determination problem is, not surprisingly, both experimentally and computationally very challenging. X-ray crystallography and nuclear magnetic resonance are two established approaches, but these cannot recover the structure of an *individual* molecule in its *natural* environment within the cell. Electron microscopy offers alternative means to study macromolecules, molecular

complexes and supramolecular assemblies in 3D. Three approaches have been developed, electron crystallography [68], single particle analysis [61], and ET.

This review focuses on ET, which is the *only* method for 3D structural studies of *individual* assemblies and subcellular structures in their *natural* environment. The idea in ET is to recover the structure of a specimen from a series of *micrographs*, which are 2D TEM images, using principles of tomography. Since its introduction to the scientific community in 1968 [41, 85, 179], there has been tremendous progress in instrumentation and sample preparation that has resolved many of the difficulties associated with data collection. Most of the computational challenges however remain to be addressed. Established approaches for image reconstruction are direct adaptations of methods from medical imaging, used with inadequate appreciation for the finer points of mathematics. Therefore, 3D reconstructions have a resolution that is reliable only for studying cellular substructures, rather than studying individual small molecules ("small" in this context refers to protein molecules of size smaller than 200 kDa). This in turn has severely limited the usefulness of ET in life sciences.

Another rapidly growing application area for ET is 3D characterization and metrology of nano-structures in material sciences and semiconductor manufacturing. In these applications, data is often acquired using a variety of different TEM imaging modes (paragraph on p. 941), see [126, 127, 181] for a nice review. The focus in this review is on life sciences, even though we occasionally mention material science applications.

2 The Transmission Electron Microscope (TEM)

The starting point of electron microscopy can be attributed to Hans Busch who laid the theoretical basis for the electron microscope and designed the first working electron lens [23, 24]. Ten years later, based on the work of Busch, Ernst Ruska constructed the first operational TEM [98]. Ruska was later awarded the 1986 Nobel Prize in physics.

The operational principle of an electron microscope is similar to a light microscope (Fig. 1). A source emits electrons that are accelerated (typically by 200 kV) on their way to the specimen. A condenser system controls how the specimen is illuminated. After passing through the electron–transparent specimen, an electron optical system collects the electrons of interest and directs them onto the image plane, where an intensity is generated. This is detected and converted to a gray-scale image (micrograph) by means of a detector system. The beam electrons scatter elastically and/or inelastically against the atoms in the specimen. This interaction can generate a range of secondary signals whose analysis reveal different properties of the specimen.

A modern TEM can be operated under different settings (*imaging modes*) to image some of the aforementioned signals. Each imaging mode is determined by (1) how the specimen is illuminated, (2) which electrons that are selected to take part

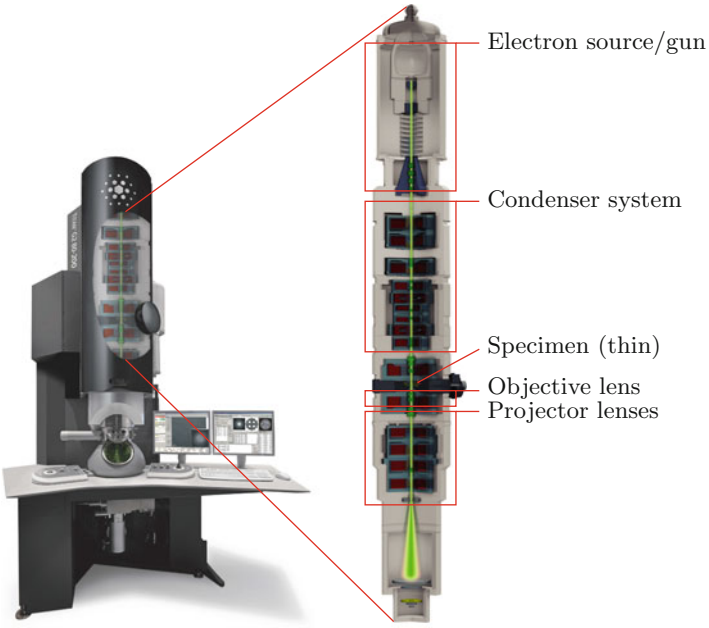


Fig. 1 A modern TEM with its column exposed. The electron gun with the source is at the *top*. Electrons travel downwards, passing through the condenser system before they scatter against the specimen. Scattered electrons pass through the optics (objective and projector lenses) and form an image at the bottom (image courtesy of FEI)

in the image formation and where they are registered, and (3) the type of detector. As an example, dark-field mode uses only the deflected electrons for image formation, whereas in bright-field mode only the undeflected electrons take part in the image formation.

ET in life sciences is mostly based on *conventional bright-field* TEM imaging. The specimen is here illuminated by a parallel beam of incident electrons (uniform illumination mode), and transmitted electrons that pass through the aperture in the back focal plane of the objective lens form an image in the image plane. The aperture is centered (bright-field mode), so only electrons that scatter with very small angle take part in the image formation.

TEM Imaging Modes

The specimen can be illuminated by focusing the electrons onto a spot (scanning mode), which leads to Scanning Transmission Electron Microscope (STEM). Next, the objective aperture can be placed as to block the direct beam of un-scattered electrons, while one or more diffracted beams are allowed to pass (dark-field mode). Finally, instead of registering the image in the image plane, it can be registered in the back-focal plane of the objective lens (diffraction contrast mode), which corresponds to imaging the diffraction pattern of the specimen.

A common mode is selected area electron diffraction that uses uniform illumination with diffraction contrast imaging, suitable for investigating spatial variation of diffraction properties in a (crystalline) specimen. Convergent-beam electron diffraction is the corresponding set-up based on scanning mode illumination. Another mode is High-Angle Annular Dark-Field (HAADF) that uses scanning mode illumination and micrographs are acquired in dark-field mode. The electrons that take part in image formation are incoherent elastically scattered electrons, so contrast depends strongly on the average atomic number of the scatterer ("Z"-contrast). Finally we mention analytical TEM that uses energy resolving detectors, Energy-Dispersive X-ray Spectroscopy (EDS) or Electron Energy Loss Spectroscopy (EELS), combined with scanning mode for illumination. This allows one to acquire a "chemical map" of the specimen. An EDS spectrometer does this by measuring the X-rays emitted from the specimen, whereas an EELS spectrometer measures those transmitted electrons that undergo a pre-specified energy loss.

The above modes are mostly used in material sciences and semiconductor industry, even though there are usages in life sciences. The interested reader may consult [63] and [7, chapter 3] for a more detailed account.

Sample Preparation

Samples to be imaged in an electron microscope generally require processing to produce a suitable specimen. Sample preparation techniques depend on the sample, the information one seeks to image, and the type of microscope. A mathematician may safely ignore most issues related to sample preparation, but some aspects of it that do have algorithmic implications. Before we discuss these (in paragraph on p. 943), we provide a very brief overview of sample preparation.

Overview

In general, all sample preparation techniques include some kind of fixation that aims at solidifying the specimen. This is needed because electron microscopy imaging is performed under high vacuum conditions. Next, thinning might be required to ensure that the specimen is thin enough so that electrons can pass through. Both these steps are done while preserving as much as possible of the structural integrity of the specimen. Finally, a user might choose to add substrates into the specimen in order to increase contrast (contrast enhancement). A wide range of techniques are now available, see [7, 8] for a comprehensive review.

In life sciences, the preferred fixation technique is by freezing (cryo-electron microscopy) [148, section 3.1]. If the specimen is liquid, or thin enough ($<10\ \mu m$), then one can use plunge-freezing. Here the specimen is applied to a support and then dropped into a cryogen (such as liquid ethane at less than $-160\ ^\circ C$) where sufficiently fast cooling speeds are reached that prevent the formation of crystalline ice. This solidifies the specimen with minimal artifacts. It is the preferred cryo-fixation method for in vitro specimens, i.e., specimens that contain free particles (proteins and macromolecular assemblies) in an aqueous solution. Samples that

are too thick for plunge-freezing are treated with high-pressure freezing, which combines rapid temperature decrease with the application of high pressure that lowers the melting point of water. It allows vitrification of specimens up to 200 μm in thickness.

Relevance to Mathematics

Let us first consider usage of contrast enhancement, like in negative staining. The high atomic numbers for contrast agents improve the signal-to-noise properties of the micrographs, so noise becomes less of a problem. On the other hand, contrast agents have difficulties reaching the interior of the molecular assemblies, so micrographs show outlines of the molecules rather than their internal structure. Therefore, contrast enhancement is reserved for low- or medium-resolution TEM imaging. In the context of ET, we can in the forward model often disregard phase contrast (section "Forward Operator for Amplitude Contrast Only"). Next, since focus is on determining the 3D shape of the sub-structures within the specimen, rather than their internal structure, 3D reconstruction methods should preserve edge information whereas smooth gray scale variations are of less interest. The impact of local data (p. 974) can also be quite severe in this setting (section on p. 985), so if possible, one should use reconstruction methods that are less sensitive to such issues.

Cryo-fixated unstained specimens are on the other hand thin and weakly scattering. ET of High-Resolution TEM (HRTEM) imaging data from such specimens must model phase contrast. Specimens that contain particles in aqueous buffer (in vitro specimens) are fixated by plunge freezing. These are imaged using a very low electron dose, so micrographs are very noisy and 3D reconstruction methods must efficiently utilize whatever a priori information that is available, like regularity or sparsity adapted to this setting. Impact from local data is on the other hand less severe since most of the specimen is vitrified aqueous buffer.

Specimens that are derived from tissue (in situ specimens) typically result in micrographs with much better signal-to-noise ratio (although not as good as negatively stained specimens). The background will have significant structure, so regularity assumptions, like sparsity, will have to be quite different from those used for in vitro specimens. The impact from local data is also more severe as compared to imaging data from in vitro specimens.

3 Basic Notation and Definitions

$\mathbb{R}$ and $\mathbb{C}$ denote the fields of real and complex numbers, $\mathbb{R}_+$ and $\mathbb{C}_+$ are positive numbers (for complex numbers, this means positive real and imaginary parts), and $\mathbb{R}^n$ denotes the n-dimensional vector space over $\mathbb{R}$. Next, for $\boldsymbol{x} \in \mathbb{R}^3$ we let $\boldsymbol{x}^\perp$ denote the hyperplane in $\mathbb{R}^3$ that is orthogonal to $\boldsymbol{x}$, i.e., $\boldsymbol{x}^\perp := \{\boldsymbol{y} \in \mathbb{R}^3 : \boldsymbol{x} \cdot \boldsymbol{y} = 0\}$. $S^2 \subset \mathbb{R}^3$ is the 3D sphere, and for $\boldsymbol{\omega} \in S^2$ and $p \in \mathbb{R}$, we let $p\boldsymbol{\omega} + \boldsymbol{\omega}^\perp$ denote the hyperplane of points $\boldsymbol{x} = p\boldsymbol{\omega} + \boldsymbol{y}$ where $\boldsymbol{y} \in \boldsymbol{\omega}^\perp$.

Next, $\mathcal{F}$ and $*$ denote the Fourier transform and convolution of functions defined on $\mathbb{R}^n$. Furthermore, given a function $U : \mathbb{R}^3 \to \mathbb{C}$ that decreases sufficiently fast at infinity, say $U \in \mathscr{S}(\Omega, \mathbb{C})$, its *ray transform* $\mathcal{P}(U)$ (also called projection in the ET community) is defined as

$$\mathcal{P}(U)(\boldsymbol{\omega}, \boldsymbol{x}) := \int_{-\infty}^{\infty} U(\boldsymbol{x} + s\boldsymbol{\omega}) \, \mathrm{d}s \quad \text{for } \boldsymbol{\omega} \in S^2 \text{ and } \boldsymbol{x} \in \boldsymbol{\omega}^{\perp}. \tag{1}$$

More precise mathematical conditions that are required for the existence of the ray transform are stated in [122, section 3.5]. Finally, following [136, eqs. (2.29) and (2.30)], for fixed $\boldsymbol{\omega} \in S^2$ we let $\mathcal{F}_{\boldsymbol{\omega}^{\perp}}$ and $\circledast_{\boldsymbol{\omega}^{\perp}}$ denote the corresponding *Fourier transform* and *convolution on the (two-dimensional) hyperplane* $\boldsymbol{\omega}^{\perp}$:

$$\mathcal{F}_{\boldsymbol{\omega}^{\perp}}[U](\boldsymbol{\xi}) := \int_{\boldsymbol{\omega}^{\perp}} U(\boldsymbol{x}) e^{-i\boldsymbol{x}\cdot\boldsymbol{\xi}} \, \mathrm{d}\boldsymbol{x} \quad \text{for } \boldsymbol{\xi} \in \boldsymbol{\omega}^{\perp},$$

$$(U \underset{\boldsymbol{\omega}^{\perp}}{\circledast} H)(\boldsymbol{x}) := \int_{\boldsymbol{\omega}^{\perp}} U(\boldsymbol{x} - \boldsymbol{y}) H(\boldsymbol{y}) \, \mathrm{d}\boldsymbol{y} \quad \text{for } \boldsymbol{x} \in \boldsymbol{\omega}^{\perp}.$$

4 The Forward Model

This section gives an overview of TEM image formation models (forward models) relevant for ET. It includes a short derivation of various models with discussions on their validity and computational feasibility. Much of the material is therefore physics related, so a reader primarily interested in image reconstruction could skip this part and directly consult (57) (or section "Forward Operator for Combined Phase and Amplitude Contrast" for a slightly more detailed summary).

Basic Assumptions

We only consider conventional bright-field TEM imaging of amorphous specimens, since this is the prevalent imaging mode for ET in life sciences. The starting point is to assume that the imaging electron and the specimen form a closed system, i.e., there is no interaction with the environment. Next, we also assume that successive imaging electrons can be treated independently and any interaction between them can be neglected (independent electron assumption). This holds under typical TEM imaging conditions in ET where the mean separation between two successive electrons is much larger than the specimen thickness and the length of the electron wave packet [63, p. 85]. Hence, wave mechanical notions, like "interference," refer to the wave crests of an *individual* electron (i.e., self-interference).

Bearing in mind these two basic assumptions, an ideal model would be based on solving the Schrödinger equation for the electron microscope (including the specimen) as a whole. This is clearly unfeasible, so the first step is to separate the problem into four parts: (1) illumination, (2) electron–specimen interaction,

(3) optics, and (4) detection. A great number of approaches have been developed for modeling each of the above parts.

Contrast Mechanisms

Models for TEM imaging fall into two categories depending on how image contrast is related to electron–specimen interaction and optics.

Amplitude contrast models (section "Forward Operator for Amplitude Contrast Only") assume all contrast variations in a micrograph are due to the removal of electrons from then beam (*amplitude contrast*). This is adequate at medium and low resolution. At higher resolution, like in HRTEM where image features of interest are smaller than the coherence length of the electron, one also needs to account for the phase shift that an electron undergoes as it scatters against the specimen (*phase contrast*). *Combined phase and amplitude contrast models*, that are briefly described below, account for both the amplitude and phase contrast.

The starting point for the combined phase and amplitude contrast model is to assume perfect coherent imaging where incident electron waves are monochromatic plane waves (uniform coherent illumination) and electrons scatter only elastically. This leads to a stationary scattering problem where the specimen remains in the same quantum state with scattering properties fully described by its electrostatic potential. The treatment of the electron is quantum mechanical. The electron–specimen interaction is then modeled by the scalar Schrödinger equation and the picture is completed by adding a description of the effects of the optics and the detector of the TEM, both modeled as convolution operators. Inelastic scattering and incoherent illumination introduce partial incoherence, so the basic assumption of perfect coherent imaging must be relaxed. Inelastic scattering can be accounted for within the coherent framework by introducing an imaginary part to the scattering potential (absorption potential). Incoherence from the illumination is usually accounted for by modifying the convolution kernel that models the optics. Finally, linearizing the scattering model and the intensity yields an explicit linear relation between the measured intensity and the scattering potential.

Illumination

This is where incident electrons are generated (source/gun) and controlled (condenser).

Basic Components

Commercially available TEMs either use a thermionic or a field-emission (electron) source. The first produces electrons by heating the source, whereas the second produces electrons by applying a large electric potential between an anode and the source. The performance of the source is characterized by its brightness, temporal coherency, energy spread, spatial coherency, and stability. The brightness (the current density per unit solid angle of the source) is the most important parameter

and it influences the resolution, contrast, and signal-to-noise capabilities of the microscope. A source with smaller size gives higher brightness and better spatial coherency, but is also less stable.

The source is incorporated into a gun that can control and focus the electrons. According to the type of the source, we have either thermionic gun or field-emission gun. The size of the first one is bigger, so it can illuminate large areas on the specimen at relatively low image magnification (<50–100,000×) without losing current density. The field emission gun has become the preferable electron source for HRTEM imaging (for both uniform and scanning mode illumination) due to its high brightness and small source size. More information can be found in standard text books on electron microscopy, see, e.g., [80, sections 43–50].

After leaving the gun, the electrons pass through a condenser system that typically consists of two condenser lenses and an aperture. This is the second most important part of the TEM, the objective lens (section "Optics") being the most important, since it determines how the specimen is illuminated. One can illuminate the specimen with a wide collimated parallel beam (uniform illumination) or with a parallel narrow beam (scanning mode illumination).

Model for Illumination

Since we consider conventional TEM, micrographs are acquired using uniform illumination mode. Furthermore, we will not directly model the source, gun, or condenser. Instead, as a starting point, we assume that an incident electron leaving the condenser is a monochromatic plane wave traveling along a fixed direction $\boldsymbol{\omega} \in S^2$ that is parallel to the TEM optical axis:

$$\Psi(\boldsymbol{x}, t) = \exp(-itE/\hbar)\exp(ik\boldsymbol{x}\cdot\boldsymbol{\omega}).$$

Here, E is the energy of the electron, $\hbar$ is the reduced Planck constant, and k is the wave number of the electron, so $k = 2\pi/\lambda$ where λ is the wavelength.

In reality, however, the illumination is only partially coherent. The effective size at the electron source is small but not point-like, and the energy spread is narrow but still appreciable. A perfectly coherent source would require vanishingly small source and negligible energy spread. The effects of such partial coherence is usually modeled by a perturbation procedure and incorporated into the TEM optics PSF, as outlined in paragraph on p. 961.

Electron–Specimen Interaction

This is the model for the scattering of an electron against the atoms in the specimen. In the general setting, it is a fairly complex process involving a variety of phenomena depending on the energy of the electron, the atomic composition of the specimen, and the type of chemical bindings between the atoms in the specimen. An accurate

and complete model for all electron scattering phenomena, that includes both quantum and relativistic aspects, requires the theory of quantum electrodynamics. A slightly less complete framework is based on the time-dependent multi-body Schrödinger equation where one approximately accounts for relativistic effects. Here, in the full model, the specimen is a quantum-mechanical many-body system whose scattering properties are accounted for by the corresponding Hamiltonian. This leads to a computationally unfeasible model, so we need further simplifications. One is to considering elastic scattering.

Elastic Scattering

Elastic scattering refers to the case when there is no (or negligible) transfer or energy from the electron to the specimen. This allows us to treat the electron–specimen interaction as a two-body problem where the specimen is described by its electrostatic potential and the electron is represented by its wave function $\Psi: \mathbb{R}^3 \times \mathbb{R} \to \mathbb{C}$, and the aforementioned multi-body Schrödinger equation for the system simplifies to a one-body Schrödinger equation [146, section 1.5]:

$$i\hbar \frac{\partial \Psi}{\partial t}(\boldsymbol{x}, t) = \frac{1}{2m}\Big[-i\hbar \nabla + e\boldsymbol{A}(\boldsymbol{x})\Big]^2 \Psi(\boldsymbol{x}, t) - eU(\boldsymbol{x})\Psi(\boldsymbol{x}, t). \tag{2}$$

In the above, m is the rest mass of the electron, e the absolute value of the elementary charge, $U: \mathbb{R}^3 \to \mathbb{R}^3$ is the electrostatic potential, and $\boldsymbol{A}: \mathbb{R}^3 \to \mathbb{R}^3$ is the magnetic vector potential for the stationary electromagnetic field surrounding the system.

Stationary Model

Since the electrostatic potential U in (2) does not depend on time, it is sufficient to consider standing wave solutions to (2):

$$\Psi(\boldsymbol{x}, t) = \exp(-itE/\hbar)\psi(\boldsymbol{x}). \tag{3}$$

In the above, $\psi: \mathbb{R}^3 \to \mathbb{C}$ is the complex valued space dependent part of Ψ, henceforth called the *electron wave*. Inserting (3) into (2), and canceling the common factor $\exp(-itE/\hbar)$ that appears throughout, results in a partial differential equation (4a) (stationary Schrödinger equation) for the electron wave ψ. Furthermore, in TEM imaging the specimen is typically "slab-like," meaning that its extension along the TEM optical axis is much smaller than its extension in the plane orthogonal to the optical axis. A natural model is therefore to assume U is supported in an infinite slab $\Omega \subset \mathbb{R}^3$ that is bounded by two parallel hyperplanes Γ_{in} and Γ_{out}.

To summarize, the electron–specimen interaction is modeled by the stationary Schrödinger equation with a Dirichlet boundary condition on Γ_{in} given by the incident electron wave and a boundary condition at infinity (Sommerfeld radiation condition):

$$\begin{cases} E\psi(\boldsymbol{x}) = \dfrac{1}{2m}\Big[-i\hbar\nabla + e\boldsymbol{A}(\boldsymbol{x})\Big]^2\psi - eU(\boldsymbol{x})\psi(\boldsymbol{x}) \quad \text{for } \boldsymbol{x} \in \Omega, & \text{(4a)} \\ \psi(\boldsymbol{x}) = \exp(ik\boldsymbol{x}\cdot\boldsymbol{\omega}) \quad \text{for } \boldsymbol{x} \in \Gamma_{\text{in}}. & \text{(4b)} \\ \lim\limits_{r\to\infty} r\Big[\mathbf{n}_r(\boldsymbol{x})\cdot\nabla\psi_{\text{sc}}(\boldsymbol{x}) - ik\psi_{\text{sc}}(\boldsymbol{x})\Big] = 0 \quad \text{for } |\boldsymbol{x}| = r. & \text{(4c)} \end{cases}$$

Here, ψ_{sc} in (4c) is the scattered part of ψ, *i.e.*, $\psi(\boldsymbol{x}) = \exp(ik\boldsymbol{x}\cdot\boldsymbol{\omega}) + \psi_{\text{sc}}(\boldsymbol{x})$ and $\mathbf{n}_r(\boldsymbol{x})$ is the outwards normal at $\boldsymbol{x}$ to the sphere with radii r. Note also that Ω is not necessarily orthogonal to the TEM optical axis.

A further simplification is when scattering takes place in a "field-free" region, i.e., curl $\boldsymbol{A} = \boldsymbol{B} \equiv \mathbf{0}$ in Ω. Due to the invariance under Gauge transformations, this implies that $\boldsymbol{A} \equiv \mathbf{0}$ in Ω. Combined with $E = \hbar^2k^2/2m$, transforms (4a) into an equation of Helmholtz type:

$$\big(\triangle + k^2\big)\psi(\boldsymbol{x}) = -\frac{2m}{\hbar^2}eU(\boldsymbol{x})\psi(\boldsymbol{x}) \quad \text{for } \boldsymbol{x} \in \Omega. \tag{5}$$

The Scattering Operator

The *scattering operator* $\mathcal{T}_{\boldsymbol{\omega}}^{\text{sc}}$ maps a potential U to the electron wave on the specimen exit plane Γ_{out}, i.e.,

$$\mathcal{T}_{\boldsymbol{\omega}}^{\text{sc}}(U) := \psi\big|_{\Gamma_{\text{out}}} \quad \text{where } \psi \text{ solves (4a)–(4c).} \tag{6}$$

Uniqueness and Stability

The scattering operator in (6) is well defined only when (4a)–(4c) has a unique solution. This follows from existence and uniqueness for the magnetic Schrödinger equation (4a) in a half space [149, theorem 2.16].

Remark 1. Formally, the uniqueness result [149, theorem 2.16] is for the case when $\boldsymbol{\omega}$ is orthogonal to Γ_{out}. The proof however extends directly to the case when $\psi = \psi_0$ on Γ_{in} where $\psi_0 \in \mathscr{L}_c^2(H, \mathbb{C})$.

Remark 2. Standard uniqueness results for the direct scattering problem, like in [36, p. 16], assume Ω is bounded and Dirichlet condition is on all of the boundary of Ω. In our scattering problem (4a)–(4c), Ω is unbounded and we only have Dirichlet conditions on a part of the boundary.

Another important aspect is stability for the direct scattering problem (4a)–(4c), which together with uniqueness implies well-posedness. The author is unaware of any stability results that are directly applicable to (4a)–(4c). On the other hand, [87] provides analytical evidence of increasing stability of the Cauchy problem for the Helmholtz equation (5) for increasing wavenumber k. Since TEM imaging is associated with high wave numbers, it is reasonable to conjecture that (4a)–(4c) has good stability properties.

Relativistic Corrections

The Schrödinger equation (2) is valid only for non-relativistic electron motion. In practice, this means kinetic energies should not exceed 10 keV. In TEM imaging, electron energies are between 100 and 300 keV, so one must consider relativistic wave equations.

A relativistically correct framework is offered by Dirac's equation that accurately models electron interaction and propagation in practically all electron optical systems. The computational complexity associated with using Dirac's equation has however limited its usefulness. An alternative approach is to use relativistically corrected quantities in (2) [146, section 1.9]. This is justified from a pragmatic point of view if since these corrections provide a model that has a good agreement with experimental results.

Remark 3. The Pauli approximation, which holds for TEM imaging in ET, allows one to reduce Dirac's equation to a scalar relativistic wave equation of the same type as (2) [79, section 56.3]. This introduces relativistic corrections into (2) in a way that has stronger theoretical justification.

Inelastic Scattering

Inelastic electron scattering is characterized by transfer of energy from the incident electron to the specimen, causing the latter to change state. For many TEM imaging modes, including conventional bright-field TEM imaging, inelastic scattering manifests itself as a blurring superimposed onto the image generated by the elastically scattered electrons [100]. Therefore, if possible, energy filters are used to remove inelastically scattered electrons. This however does not completely resolve the issue of inelastic scattering, since energy filtered micrographs have contributions from inelastic scattering events associated with energy losses too small to be filtered ($<$1–5 eV) [45].

The Absorption Potential

In a rigorous quantum mechanical model for inelastic electron scattering, the quantum state of the specimen is changed. Formally, this requires a time-dependent scattering model with a wave function and an interaction potential that depends on all internal degrees of freedom of the system (specimen and electron). This is clearly computationally unfeasible. An alternative approach, that is frequently used for modeling TEM imaging of amorphous specimens at 100–300 keV, is to approximately account for inelastic scattering within the framework offered by the one-body Schrödinger equation (2).

The idea is to split the aforementioned interaction potential into two parts, one time-dependent and the other time-independent [146, p. 3]. The time-independent part is obtained by taking the time-average of the interaction potential and it models elastic scattering, whereas the time-dependent part models inelastic scattering. Inelastic scattering can now be accounted for using a first-order perturbation analysis w.r.t. time as it tends to infinity. This gives rise to the appearance of an additive imaginary part of the time-averaged potential (*absorption (optical)*

potential) [146, p. 3], thus formally replacing the real valued function electrostatic potential in (2) with a complex valued one, i.e., $U = U_{\text{re}} + iU_{\text{im}}$ where $U_{\text{re}}: \mathbb{R}^3 \to \mathbb{R}$ is the electrostatic potential and $U_{\text{im}}: \mathbb{R}^3 \to \mathbb{R}$ is the aforementioned absorption potential.

Interpretation of the Absorption Potential

The electrostatic potential U_{re} models the phase shift that an electron undergoes as it scatters against the specimen (phase contrast), whereas the absorption potential U_{im} models the decrease in the flux, due to inelastic scattering, of the non-scattered and elastically scattered electrons (see [146, p. 5 and section 1.10] for a definition of the notion of flux in wave mechanics). Hence, the absorption potential accounts for amplitude contrast due to inelastic scattering. A more detailed quantum mechanical interpretation of the absorption potential is provided in [128][146, section 1.8 and chapter 13] [176, section 5.9].

Now, any computational treatment of TEM imaging data where phase contrast is important, like HRTEM data, has to address the phase problem (section "Phase Retrieval"). In many applications, like for ET in life sciences, the specimen is viewed only once from a fixed direction. Resolving the phase problem in such cases requires one to utilize further knowledge about the complex valued potential U. Now, in the context of TEM imaging in life sciences, we argue below that U_{re} and U_{im} can be considered approximately proportional. This is enough for resolving the phase problem in the context of ET.

Most inelastic scattering events in conventional bright-field TEM imaging are either due to plasmon excitation or atomic inner-shell ionization (atom core losses), where the former is considered more dominant. Next, inelastic scattering events are more frequent for scattering against light atoms, they are almost always incoherent and associated with low scattering angles, and their likelihood increases with specimen thickness. To summarize, plasmon excitation from light atoms constitute the main contribution to the inelastic scattering that damps the interpretable phase contrast signal. For TEM imaging of amorphous biological specimens, the light atoms mostly make up the background, which for cryo-fixated specimens would be the embedding medium, i.e., the main contribution to inelastic scattering comes from the vitrified aqueous buffer. Now, the imaginary part of the potential associated with plasmon excitation for an amorphous embedding medium, like vitreous ice, is expressible by the inelastic mean free path Λ_{inel}:

$$U_{\text{im}} = \frac{1}{2\sigma\Lambda_{\text{inel}}} \quad \text{where } \sigma := \frac{me}{k\hbar^2}.$$

This forms the basis for a common assumption in ET, namely that the real and imaginary parts U_{re} and U_{im} of U in (2) are proportional to the path length of the electron through the specimen. Stated more precisely, there is a function $Q: S^2 \to \mathbb{R}$ such that

$$\mathcal{P}(U_{\text{im}})(\boldsymbol{\omega}, \boldsymbol{x}) \approx Q(\boldsymbol{\omega})\mathcal{P}(U_{\text{re}})(\boldsymbol{\omega}, \boldsymbol{x}) \quad \text{for } \boldsymbol{x} \in \boldsymbol{\omega}^{\perp} \text{ and } \boldsymbol{\omega} \in S^2. \tag{7}$$

Here, $\mathcal{P}$ is the ray transform in (1) and $\boldsymbol{\omega}$ is the direction of the incident electron. A common further simplification is to let Q be a constant (amplitude contrast ratio), which then can be estimated from data (section on p. 985).

Remark 4. Assumptions of the type in (7) that relate the real and imaginary parts of U are known as the "Pagnini approach" in X-ray phase contrast tomography community [139].

More Accurate Quantum Mechanical Models

A fundamental difference between modeling inelastic, as opposed to elastic, scattering is that the former requires knowledge of the local scattering and excitation properties of the specimen. The scattering properties of the specimen are no longer fully described by a single real-valued function (the electrostatic potential). In general one would need to specify the elements of the transition matrix that encode the probability that a specimen undergoes a specific transition from one excited state to another. This situation can be somewhat simplified, under certain conditions it is, e.g., possible to express the inelastic cross section without explicit knowledge of these excited states [79, section 68.6]. It is also clear that a computationally feasible approach cannot be based on migrating a wave function that depends on the spatial location of all subatomic particles in the specimen.

A theoretically stringent, yet computationally feasible, framework is based on migrating the mutual coherence function instead of the wave function, the former being the least amount of information needed to describe interference. The scattering properties of the specimen are then modeled by the mixed dynamic form factor that accounts for the interference in Fourier space of different scattered partial electron waves by their mutual coherence [160]. Approaches based on this framework are, e.g., described in [107, 129, 183]. Unfortunately, these approaches are still computationally unfeasible in the setting when the specimen is amorphous and/or thicker than a few atomic layers, thus limiting their usefulness in ET in life sciences.

Importance of Inelastic Scattering

The absorption potential model outlined above is adequate for modeling most conventional bright-field HRTEM imaging problems with amorphous specimens. Combined with (7) and geometric optics approximations (section on p. 954), one gets a computationally feasible model for the electron–specimen interaction in the context of ET.

The validity of this image formation model is less clear for TEM imaging of specimens in material sciences. As an example, the absorption potential approach is not sufficiently accurate when imaging crystalline specimens in diffraction mode and most of [146, chapters 1 and 7–8] is devoted to presenting various models that are more accurate. Another situation that requires more accurate models of inelastic scattering is when specimens are imaged using lower voltages. The motivation is from material sciences where thin crystalline specimens undergo radiation damage when imaged using acceleration voltages beyond 100 kV. For low accelerating voltages ($\sim$20 kV), most specimens act as strong scatterers. This requires accurate

modeling of inelastic scattering, especially when atoms are light as in biological specimens. In this context, models based on the mutual coherence function are the most promising [107].

Properties of the Scattering Potential

Let $\mathscr{X}$ denote the set of complex valued functions on $\mathbb{R}^3$ that can serve as a potential $U:\mathbb{R}^3 \to \mathbb{C}$ in (2). Elements in $\mathscr{X}$ have to posses certain regularity, both from a mathematical and a physics point-of-view. One such regularity property motivated from the physics is that U fulfills the Rollnick condition, which is the case when

$$\mathscr{X} \subset \mathscr{L}^1(\Omega, \mathbb{C}) \bigcap \mathscr{L}^2(\Omega, \mathbb{C}).$$

The Rollnick condition guarantees that the corresponding Hamiltonian is self-adjoint and one has asymptotic completeness. The latter means that no matter how the wave-packet is made, particles go far apart from each other when time becomes large, so the probability of finding them together in an arbitrary finite region of space goes to zero. This allows one to go from the time-dependent Schrödinger description to the time-independent description characterized by solutions of type (3) [31, p. 133]. See also [195] for the other regularity properties associated with physics.

Next, we claim that $U:\mathbb{R}^3 \to \mathbb{C}_+$ whenever $U \in \mathscr{X}$. The positivity of the imaginary part follows from the observation that once there is an inelastic scattering event, that partial wave can never be considered as un-scattered or inelastically scattered (for that it would have to gain energy from the specimen). Next, the real part is the electrostatic potential that, at sub-atomic scale, might actually change sign. It is made up of two contributions, one "atomic" given by the superposition of atomic potentials as if each atom was in isolation, and one from the charge redistribution due to the solvent, ions, and molecular interactions. In the context of ET, the isolated atom superposition approximation allows one to account for the latter contributions within the "atomic" contributions. Hence, the specimen can be seen as consisting of isolated neutral atoms, each modeled as a point charge representing the nuclei surrounded by a spherical shell. The sphere represents the screening imposed by the shell electrons, so the electrostatic potential from the electrons cancels the one from the nucleus outside the sphere (the atom is neutral). Since the radius of the sphere is much larger than the radius of the nucleus (which here is a point charge), the electrostatic potential from the nucleus must dominate the electrostatic potential from the shell electrons within the sphere. Hence, the electrostatic potential is non-negative everywhere.

As a final note, one can put further regularity on elements in $\mathscr{X}$, e.g., $\mathscr{X} \subset \mathscr{S}(\Omega, \mathbb{C})$ (rapidly decreasing) and/or $\mathscr{X} \subset \mathscr{BV}(\Omega, \mathbb{C})$ (bounded variation).

Computationally Feasibility

The Multi-scale Nature of Electron Scattering

Calculating the scattered electron wave, i.e., numerically solving (4a)–(4c), is a multi-scale problem. To see this, we compare the wavelength λ of the electron to the (1) size of the region of interest, (2) the variation of the electrostatic potential, and (3) size of pixels where the intensity is formed. TEM imaging typically operates with energies between 200 and 300 kV resulting in electrons with a wavelength between 0.0025 and 0.0020 nm. In the example below, we consider the case of 200 kV TEM imaging, so $\lambda \approx 0.0025$ nm.

Size of region: When imaging biological specimens, a typical region of interest $\Omega_0 \subset \Omega$ is a rectangular box where the smallest side is the slab thickness, which typically is $\geq$50 nm. For simplicity, let Ω_0 be a cube with a side length of 50 nm $\approx 20{,}000\lambda$. A straightforward numerical discretization of (4a) would require about ten points per wavelength (in theory, two points per wavelength is enough but with limited precision arithmetic, one needs 10–20 points per wavelength to reliably solve Helmholtz type of wave equations using a finite element method [121]). In this case we would have to sample $(10 \cdot 25{,}000)^3 \approx 1.6 \cdot 10^{16}$ points. Even in single precision (four bytes per value), this would require 58,200 Tbytes of memory!

Contrast: To compare the variation of the electrostatic potential U against the wavelength λ, we introduce

$$C := \lambda^2 \frac{me}{2\pi^2\hbar^2} \max_{x \in \Omega_0} \left| U(\boldsymbol{x}) \right|.$$

For water $U(\boldsymbol{x}) \approx 4.9$ V so $C \approx 1.26 \cdot 10^{-5}$, for proteins 30 V $\leq U(\boldsymbol{x}) \leq$ 70 V, so $C \approx 1.8 \cdot 10^{-4}$, and for single colloidal gold particles (typically embedded in biological specimens as fiducial markers) $U(\boldsymbol{x}) \approx 280$ V so $C \approx 7.2 \cdot 10^{-4}$.

Size of pixels: If we image at 65,000× magnification (a rather common setting for imaging subcellular structures in biology), the effective pixels size of a typical TEM detector becomes 0.23 nm $\approx 115\ \lambda$. Hence, the intensity (square norm of the scattered wave that is a solution to (4a)–(4c)) will be averaged over 2D squares with a side length of 115 λ.

In summary, the multi-scale nature of numerically solving (4a)–(4c) requires one to consider various approximations.

Geometrical Optics Approximation

Geometrical optics is an approximate treatment of wave propagation where the wavelength is considered to be infinitesimally small (semi-classical approximation). The idea is to represent the highly oscillating solution as a product of a slowly varying amplitude function and an exponential function of a slowly varying phase multiplied by a large parameter.

The starting point is the Wentzel–Kramers–Brillouin (WKB) approximation for the stationary Schrödinger equation (4a). For a first order approximation, this method seeks an asymptotic solution of the form

$$\psi(\boldsymbol{x}) = a(\boldsymbol{x}) \exp\left(i S(\boldsymbol{x})/\hbar\right) + O(\hbar) \tag{8}$$

where the amplitude a and the phase S are smooth real-valued functions independent of $\hbar$. Substituting (8) into the stationary Schrödinger equation (4a), equating the real and imaginary parts, canceling the common phase factor $\exp(i S(\boldsymbol{x}, t)/\hbar)$ that appears throughout, and ignoring $\hbar^2$-order terms (first order WKB approximation) give us

$$\left| \nabla S(\boldsymbol{x}) + e\boldsymbol{A}(\boldsymbol{x}) \right|^2 = 2meU(\boldsymbol{x}) + 2mE \tag{9}$$

$$\operatorname{div}\Big[a(\cdot)^2\big(\nabla S(\cdot) + e\boldsymbol{A}(\cdot)\big)\Big](\boldsymbol{x}) = 0. \tag{10}$$

Thus, within the first order WKB approximation, (4a) is solved by (8) in which the phase function S solves (9) (an eikonal equation of Hamilton–Jacobi type), and the amplitude term a solves (10) (transport equation).

The eikonal equation (9) can be solved by the ray tracing method [147], which in turn is based on the method of characteristics. In brief, one introduces a family of curves (rays) $s \mapsto \boldsymbol{\gamma}(s)$ which are perpendicular to the level curves (wavefronts) of S. These rays define a new coordinate system where the eikonal equation reduces to a far simpler, linear, ordinary differential equation. This ordinary differential equation for the phase S can be solved simply by integrating along the aforementioned ray, leading to

$$S\big(\boldsymbol{\gamma}(t)\big) = S\big(\boldsymbol{\gamma}(0)\big) + \int_0^t H\big(\boldsymbol{\gamma}(s)\big) - e\dot{\boldsymbol{\gamma}}(s) \cdot \boldsymbol{A}\big(\boldsymbol{\gamma}(s)\big)\, \mathrm{d}s. \tag{11}$$

In the above, $H\colon \mathbb{R}^3 \to \mathbb{R}$ is given by $H(\boldsymbol{x}) := \sqrt{2meU(\boldsymbol{x}) + 2mE}$ and the ray $\boldsymbol{\gamma}$ is given by the Lorentz equation:

$$\frac{d}{ds}\Big(H\big(\boldsymbol{\gamma}(s)\big)\dot{\boldsymbol{\gamma}}(s)\Big) = \nabla H\big(\boldsymbol{\gamma}(s)\big) + e\boldsymbol{B}\big(\boldsymbol{\gamma}(s)\big) \times \dot{\boldsymbol{\gamma}}(s) \tag{12}$$

where $\boldsymbol{B} = \operatorname{curl} \boldsymbol{A}$ is the magnetic field [79, eq. (57.11)]. Finally, combining (11) with (8) and (6) gives us

$$\mathcal{T}_{\omega}^{\mathrm{sc}}(U)(\boldsymbol{x}) \approx \psi_0(\boldsymbol{x}) \exp\left(\frac{i}{\hbar}\int_0^t H\big(\boldsymbol{\gamma}(s)\big) - \sqrt{2mE}\,\mathrm{d}s\right) \tag{13}$$

with ψ_0 denoting the incident un-scattered electron wave.

Validity of the WKB Approximation

In our specific setting, the first order WKB approximation is valid if

$$\left|\frac{\triangle\, a(\boldsymbol{x})}{a(\boldsymbol{x})}\right| \ll \frac{1}{\lambda^2}. \tag{14}$$

For TEM imaging, the above is valid for electron motion in macroscopic fields, far from their singularities. The situation for scattering in microscopic fields is less obvious. One issue is that the eikonal equation (9) can become singular (caustic formation) at points where rays intersect and the amplitude blows up. There could also be several ray trajectories joining the two points, even though this particular issue can be handled within the semiclassical limit, see, e.g., [89, theorem 1.2].

Remark 5. The expression in (13) only makes use of the first order WKB approximation. The correct semiclassical limit to (4a) contains several phases, so the asymptotic solution is a sum of functions of the form (8).

Remark 6. Gaussian beams is another high frequency asymptotic model that is closely related to geometrical optics. Here the phase is complex-valued, so there is no breakdown at caustics. The solution is still assumed to be of the form (8), but it is concentrated near a single ray of geometrical optics, see [113, 167] for further details. See also [142] for a closely related approach.

The Small Angle, Projection, and Weak Phase Object Approximations

The following approximations all aim to further simplify (13). These approximations apply in particular to TEM imaging of weakly scattering specimens, such as unstained biological specimens.

The Small Angle Approximation

If $\big|eU(\boldsymbol{x})\big| \ll E$, which typically holds for TEM imaging in life sciences, then

$$\frac{1}{\hbar}H(\boldsymbol{x}) - \sqrt{2mE} = \frac{1}{\hbar}\sqrt{2meU(\boldsymbol{x}) + 2mE} - \sqrt{2mE} \approx \frac{m}{\hbar^2 k^2} eU(\boldsymbol{x}).$$

Hence, introducing $\sigma := me/k\hbar^2$, (13) becomes

$$\mathcal{T}_{\omega}^{\mathrm{sc}}(U)(\boldsymbol{x}) \approx \psi_0(\boldsymbol{x}) \exp\left(i\sigma\int_0^t U\big(\boldsymbol{\gamma}(s)\big)\,\mathrm{d}s\right) \quad \text{for } \boldsymbol{x} = \boldsymbol{\gamma}(t) \in \Gamma_{\mathrm{out}}. \tag{15}$$

This approximation is usually referred to as the *small angle approximation*.

The Projection Approximation

Assume Ω is thin enough that we can disregard the curvature of the electron trajectories, i.e., the un-scattered electrons travel along straight lines parallel to the direction $\boldsymbol{\omega}$ of the incident plane wave ψ_0. Assuming $\boldsymbol{x} + t\boldsymbol{\omega} \in \Gamma_{\text{out}}$, the above considerations combined with (15) result in

$$\mathcal{T}^{\text{sc}}_{\boldsymbol{\omega}}(U)(\boldsymbol{x}+t\boldsymbol{\omega}) \approx \psi_0(\boldsymbol{x}+t\boldsymbol{\omega}) \exp\left(i\sigma \int_0^t U(\boldsymbol{x}+s\boldsymbol{\omega})\,\mathrm{d}s\right) \quad \text{for } \boldsymbol{x} \in \Gamma_{\text{in}}. \tag{16}$$

Furthermore, one frequently replaces $\psi_0(\boldsymbol{x} + t\boldsymbol{\omega})$ with $\psi_0(\boldsymbol{x})$, which is perfectly fine when ψ_0 is a plane wave traveling along $\boldsymbol{\omega}$ (the difference is then merely a constant pure phase factor). This is the *projection assumption* [78].

Weak Phase Object Approximation

The *weak phase object approximation* is simply based on linearizing the exponential in (16), i.e.,

$$\mathcal{T}^{\text{sc}}_{\boldsymbol{\omega}}(U)(\boldsymbol{x}+t\boldsymbol{\omega}) \approx \psi_0(\boldsymbol{x}+t\boldsymbol{\omega})\left(1+i\sigma \int_0^t U(\boldsymbol{x}+s\boldsymbol{\omega})\,\mathrm{d}s\right) \quad \text{for } \boldsymbol{x} \in \Gamma_{\text{in}}. \tag{17}$$

Other Approaches

A variety of approaches have been developed for simulating electron scattering in a TEM. Most of them, like the Bloch wave method, are only of interest for modeling TEM imaging in diffraction mode which is important in electron crystallography and in some material sciences applications. There are however approaches that are also applicable for computational treatment of conventional bright-field TEM imaging of amorphous specimens.

First Order Born Approximation

The starting point is to reformulate (4a) as an integral equation (Lippmann–Schwinger equation). The latter can be solved by means of an iterative procedure, and the first step in that procedure yields the first order Born approximation. This results in an affine model for $\mathcal{T}^{\text{sc}}_{\boldsymbol{\omega}}$, the details are given in [51, section 4.3]. Related approximations of $\mathcal{T}^{\text{sc}}_{\boldsymbol{\omega}}$ are the paraboloid approximation [90] and the thick-phase grating approximation [189].

Multi-Slice Method

Here one assumes that $\psi(\boldsymbol{x}) = \phi(\boldsymbol{x}) \exp(ik\boldsymbol{x} \cdot \boldsymbol{\omega})$ where ϕ is the slowly varying part whose sampling in real space may be spaced many wavelengths apart. Inserting into (4a) and making use of the Lippmann–Schwinger formulation result in an equation for the slowly varying part ϕ. The differential operators in this equation can now be separated into operators along $\boldsymbol{\omega}$ and operators acting in $\boldsymbol{\omega}^{\perp}$. Ignoring back-scattering yields a recursive scheme for calculating $\mathcal{T}^{\text{sc}}_{\boldsymbol{\omega}}$ on Γ_{out}, see [146, section 3.4].

Distorted Wave Born Approximation

The idea here is to express the potential as a sum of two terms, $U = U_0 + \triangle U$ where U_0 is not necessarily small and $\triangle U$ is sufficiently small that only linear terms in the expression for the scattered electron wave need to be retained. This expression for U is inserted into the Lippmann–Schwinger formulation of (4a), thereby giving rise to a series expansion [146, eq. (7.8)] where one can collect terms in $\triangle U$ and U_0. Retaining a first order correction for U_0 results in the approximate description of the scattered electron wave, see [146, section 7.2] for details.

Optics

After interacting with the specimen, electrons pass through the TEM optics as they migrate from the specimen exit plane Γ_{out} to the image plane Γ_{imag} (note that the former is not necessarily orthogonal to the optical axis whereas the latter always is). An obvious role of the optics is to magnify the image. Another equally important, but more subtle, role is related to phase contrast. The optics is necessary to make phase contrast visible, which is important for HRTEM imaging. Phase contrast would be lost if one would measure intensity data directly on Γ_{out}.

Remark 7. There are situations when one can, at least approximately, reconstruct the electron wave function from measured image intensity. Some of these techniques involve unconventional imaging modes, of which the most successful is off-axis holography using an electron biprism, see [79, section 63.2].

Thus, *any forward model that seeks to account for phase contrast must model the TEM optics*. This is especially important for imaging weakly scattering specimens, like unstained thin biological specimens.

Here we only state the model for the TEM optics and refer to [51, section 5] for its derivation. See also [80, 161, 162] and [79, sections 64–66] for a thorough description of models for various electron optical elements.

The General Setting

The optics of the TEM is the portion between the specimen exit plane Γ_{out} and the image plane Γ_{imag} (Fig. 2). Here there is no specimen, but there are strong magnetic fields that can deflect the electron beam in a desirable manner. Disregarding polarization and the precession imposed by the electron spin (which does not appreciably affect the motion of the electron), the TEM optics is fully characterized by a specification of apertures and the magnetic vector potential that corresponds to the aforementioned magnetic fields.

We define the *optics operator* as the mapping that associates an electron wave in Γ_{out} to the corresponding wave in Γ_{imag}. In the general setting, it is the solution to (4a) with $U \equiv 0$ (no specimen) and appropriate boundary conditions:

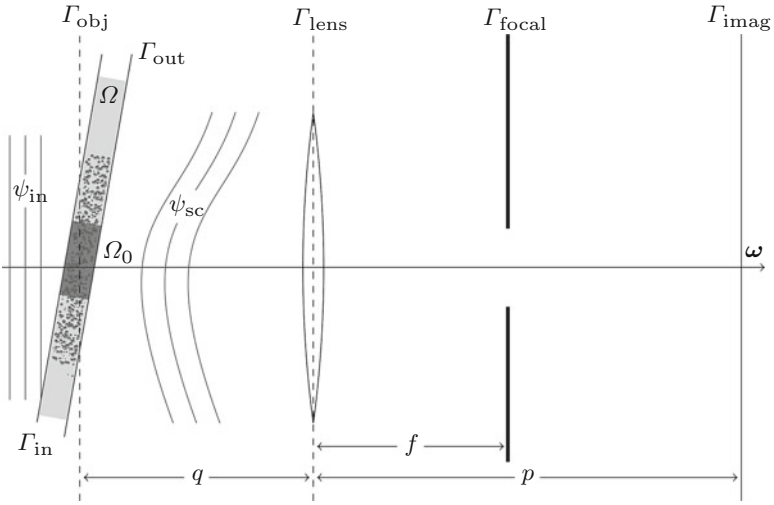


Fig. 2 The optical set-up (Ω and Ω_0 are in reality much smaller) for a single thin lens with an aperture in the focal plane (section on p. 960). The case for lens-less imaging (section on p. 960) is obtained by removing the lens and its aperture. The incident electron wave ψ_{in} (plane wave from left traveling along the optical axis $\boldsymbol{\omega}$) scatters against the atoms in the specimen characterized by $U : \Omega \to \mathbb{C}$ (scattering potential) with support in a slab Ω (*light-gray region*). The subset $\Omega_0 \subset \Omega$ (*dark-gray region*) is the region of interest. The operator $\mathcal{T}^{\text{sc}}_{\boldsymbol{\omega}}(U)$ in (6) maps ψ_{in} on Γ_{in} to ψ_{sc} on Γ_{out}. The operator $\mathcal{T}^{\text{op}}$ in (19) maps an electron wave on the object plane Γ_{obj} onto an electron wave on the image plane Γ_{imag}. Finally, $\mathcal{T}^{\text{tiltcorr}}$ defined on paragraph on p. 961 maps the electron wave on the specimen exit plane Γ_{out} onto the wave on the object plane Γ_{obj}

$$\begin{cases} E\psi = \dfrac{1}{2m}\Big[-i\hbar\nabla + e\boldsymbol{A}\Big]^2\psi \quad \text{on } \mathbb{R}^3 \setminus \Omega, \\ \psi = \psi_{\text{sp}} \quad \text{on } \Gamma_{\text{out}}, \\ \psi \text{ fulfls (4c)}. \end{cases} \tag{18}$$

Since $\Gamma_{\text{imag}} := \boldsymbol{\omega}^{\perp} + p\boldsymbol{\omega}$, the optics operator is given as

$$\mathcal{T}^{\text{op}}(\psi_{\text{sp}})(\boldsymbol{x}) = \psi(\boldsymbol{x} + p\boldsymbol{\omega}) \quad \text{for } \boldsymbol{x} \in \boldsymbol{\omega}^{\perp} \text{ and } \psi \text{ solves (18)}. \tag{19}$$

The magnetic vector potential $\boldsymbol{A}$ in (18) is given by the magnetic fields generated by the optical system and the electron wave ψ_{sp} on Γ_{out} is the scattered electron wave leaving the specimen.

Ray Optics

Many electron optical elements, including electron lenses, are modeled by considering geometric optics approximation of (18) where the electron is a point-like charged mass whose motion is governed by the laws of classical mechanics [162,

chapter 1] and [79, p. 1262]. The relation between the wave mechanical and classical formulations can in fact be seen directly from (12). This shows that the eikonal curve derived within the WKB approximation coincides with the classical electron trajectory, so the fundamental laws of geometrical electron optics are an approximate consequence of wave mechanics based on the first order WKB approximation. *In conclusion, electron lenses and their aberrations are adequately modeled using geometrical charged-particle optics. Modeling diffraction by an aperture (which is an opaque screen with a suitable opening) needs however to be based on wave mechanics.*

Remark 8. A theory of electron optics that does not make use of geometric optics approximation becomes quite complicated, see, e.g., [79, sections 59.4–59.6] [111]. Furthermore, there are many other elements of the TEM that we do not model, notably alignment coils, stigmators, and phase plates. Stigmators are used to correct for the unavoidable third-order (and possibly higher) spherical aberration of the objective lens [80, chapter 41].

The Optical Set-Up

The most important component of the optics is the *objective lens*, which provides the first stage of magnification (about 20–50 times). It is followed by a number of *projector lenses* that provide further magnification. Apertures are present at several places, e.g., in the back-focal plane of the objective lens (Fig. 1).

To model phase contrast imaging in ET, it turns out that one *can model the entire TEM optical system as a single thin lens with an aperture in its focal plane as illustrated in Fig. 2* (see Remark 9 for an motivation). Within this setting, except for the specimen exit plane Γ_{out}, all the other planes are parallel to each other and orthogonal to the optical axis, i.e., orthogonal to $\boldsymbol{\omega}$. Since the magnification of the single thin lens corresponds to the magnification M of the *entire* optical system (objective and projector lenses taken together), we get

$$M = p/q \quad \text{and} \quad 1/f = 1/p + 1/q. \tag{20}$$

Here, f is the focal length of the lens, and q, $p > 0$ are the distances from the lens to the objective and images planes Γ_{obj} and Γ_{imag}, respectively. This set-up does not correspond to a physical optical system, so one needs to set values for f, p, and q (section "Illumination and Optics Parameters").

Remark 9. The nontrivial part in replacing the TEM optical system with a the system in Fig. 2 is to argue that the thin lens in Fig. 2 has the same aberration properties as the objective lens. The motivation for this is as follows: Due to the initial magnification by the objective lens, the angular range of electron beams impinging upon the first image plane is very small relative to the angular range of electron beams entering the object lens from the specimen. Since aberrations are worse at high angles than at low angles, the lens that is most affected by aberrations will be the objective lens as it has to deal with largest range of angles in the whole

microscope. The lenses below the objective lens (projector lenses) also magnify, and as the magnification increases, the range of angles that each subsequent lens must deal with is reduced. The first lens of the projector system (sometimes called the "intermediate" or "diffraction" lens) matters a little bit, but one can forget about all the other lenses. Even though they provide a huge amount of magnification, they have virtually no influence on the final image resolution.

Lens-Less Imaging

This is the simplest model where the optics is replaced by free space propagation corresponding to the distance between the parallel object and image planes. Thus, one ignores any magnetic fields (so $\boldsymbol{A} \equiv \mathbf{0}$, see text preceding (5)), apertures, and the effect of specimen tilt (see paragraph on p. 961). Despite its simplicity, this model is much better than simply assuming the intensity is taken directly at the specimen exit plane. The reason is that free-space propagation gives time for the phase shift induced by the specimen to develop into visible contrast, thereby partly addressing the phase problem.

Stated mathematically, free-space propagation is given as the solution to (18) with $\boldsymbol{A} \equiv \mathbf{0}$ and Γ_{out} replaced by Γ_{obj}. The resulting equation can be solved under the Fresnel approximation and the solution is expressible in terms of a 2D convolution in the object plane:

$$\mathcal{T}^{\text{op}}(\psi_{\text{sp}})(\boldsymbol{x}) \approx \exp\big(ik(p-q)\big)\mathcal{T}^{\text{fsp}}_{p-q}\big(\psi_{\text{sp}}(q\boldsymbol{\omega}+\cdot)\big)(\boldsymbol{x}) \quad \text{for } \boldsymbol{x} \in \boldsymbol{\omega}^{\perp}, \tag{21}$$

where, for $d > 0$,

$$\begin{cases} \mathcal{T}^{\text{fsp}}_{d}(\phi)(\boldsymbol{x}) := \dfrac{k}{2\pi i d}\Big\{\text{PSF}^{\text{fsp}}_{d} \underset{\boldsymbol{\omega}^{\perp}}{\circledast} \phi\Big\}(\boldsymbol{x}) & \text{(22a)} \\ \text{PSF}^{\text{fsp}}_{d}(\boldsymbol{x}) := \exp\Big(i\dfrac{k}{2d}|\boldsymbol{x}|^2\Big). & \text{(22b)} \end{cases}$$

Single Thin Lens with an Aperture

Here we consider the setup in Fig. 2. The optics operator can then be modeled by a suitable combination of free-space propagation (22a), a model for a thin lens, and a model for diffraction by an aperture.

For simplicity let us disregard the correction for the specimen tilt (section "Model Refinements"). After some lengthy manipulation, shown in [51], one arrives at the following explicit expression for the optics operator:

$$\mathcal{T}^{\text{op}}(\psi_{\text{sp}})(\boldsymbol{x}) \approx \frac{\Phi(\boldsymbol{x})}{(2\pi)^2 M}\Big\{\text{PSF}^{\text{op}} \underset{\boldsymbol{\omega}^{\perp}}{\circledast} \psi_{\text{sp}}(q\boldsymbol{\omega}+\cdot)\Big\}\Big(\frac{\boldsymbol{x}}{M}\Big) \quad \text{for } \boldsymbol{x} \in \boldsymbol{\omega}^{\perp}. \tag{23}$$

Here, $M = (p-f)/f$ denotes the magnification, Φ is a pure phase factor (so $|\Phi(\boldsymbol{x})| = 1$) whose precise expression is given in [51, eq. (21)], and

$$\mathrm{PSF}^{\mathrm{op}}(\boldsymbol{x}) := \mathcal{F}_{\boldsymbol{\omega}^\perp}\left[A_\Sigma\left(\frac{f}{k}\cdot + f\boldsymbol{\omega}\right)\exp\left(iW\left(|\cdot|^2\right)\right)\right](\boldsymbol{x}) \quad \text{for } \boldsymbol{x} \in \boldsymbol{\omega}^\perp. \tag{24}$$

The Fourier transform of $\mathrm{PSF}^{\mathrm{op}}$, which by above has a closed form expression, is usually called the Contrast Transfer Function (CTF). The function $A_\Sigma : \boldsymbol{\omega}^\perp \to \mathbb{R}$ is the characteristic function for the aperture $\Sigma \subset \Gamma_{\mathrm{focal}}$ (pupil function) and $W : S^2 \times \mathbb{R} \to \mathbb{R}$ encodes the phase shift associated with imperfections on the optics (e.g., defocus, aberrations). The latter has an explicit expression [79, eq. (65.29)], which in the context of ET can be significantly simplified since terms accounting for astigmatism and higher order aberrations can be ignored [79, eq. (65.30)]:

$$W(t) := -\frac{1}{4k}t\left(\frac{C_s}{k^2}t - 2\triangle z\right) \quad \text{for } t \in \mathbb{R}. \tag{25}$$

The constant C_s is the *third-order spherical aberration of the lens* and $\triangle z$ is the *defocus* (the deviation from the focal length and error in the positioning of the specimen plane).

Remark 10. The convolution with the optics PSF in (24) is actually taken against $\psi(-q\boldsymbol{\omega} - \cdot)$. The minus sign in the argument of ψ comes from the fact that a single thin lens reverses the image. The actual optics consists of several lenses and microscope manufacturers do compensate for such effects, so we may disregard these minus signs.

Model Refinements

Specimen Tilt

When the specimen is tilted, i.e., when the slab Ω is not orthogonal to the optical axis, then one needs to migrate the electron wave from the specimen exit plane Γ_{out} to the object plane Γ_{obj}. We define $\mathcal{T}^{\mathrm{tiltcorr}}$ as the operator that maps an electron wave on the specimen exit plane to the corresponding electron wave on Γ_{obj}. Since Γ_{out} and Γ_{obj} are not parallel, this amounts to a free-space propagation between two non-parallel planes.

Free-space propagation between non-parallel planes has been worked out for electromagnetic waves in [6, 125], but the methods are also applicable to TEM imaging. Approaches in TEM imaging are based on modifying the point spread function for the lens operator ("tilted CTF"), see, e.g., [78, section 3.3] and [184, 185].

Partially Incoherent Illumination

For perfect coherent illumination, the incident electron beam should have no energy spread (perfect temporal coherence) and electrons are emitted from a vanishingly small source (perfect spatial coherence). In practice, energy spread may reach a

few eV depending on the type of electron source and the finite source size spreads the directions of arrival of the incident electrons at the specimen input plane Γ_{in}.

The standard approach for accounting these effects is to represent each of them by a damping in the Fourier space of the optics PSF. The details are given [79, sections 66.2–66.3], below we simply state the resulting modified optics PSF that replaces the one in (24):

$$\text{PSF}^{\text{op}}(\boldsymbol{x}) := \mathcal{F}_{\boldsymbol{\omega}^\perp}\left[A_\Sigma\left(\frac{f}{k}\cdot + f\boldsymbol{\omega}\right) \exp\left(iW\left(|\cdot|^2\right)\right) E_{\text{spr}}\left(|\cdot|^2\right) E_{\text{size}}\left(|\cdot|^2\right)\right](\boldsymbol{x}).$$

The functions E_{spr} and E_{size} are the envelope functions that accounts for the energy spread and the extension of the source, respectively:

$$E_{\text{spr}}(t) := \exp\left(-\frac{\pi^2}{4f^4}\nu_m^2 C_c'^2 t^2\right)$$
$$E_{\text{size}}(t) := \exp\left(-\frac{1}{4}\frac{k^2}{f^2}\alpha_c^2 t\left(\frac{C_s}{f^2}t - \triangle z\right)^2\right).$$

In the above, ν_m denotes the mean energy spread,

$$C_c' := \frac{1 + 2\epsilon U_{\text{acc}}}{U_{\text{acc}}(1 + \epsilon U_{\text{acc}})} C_c \quad \text{where} \quad \epsilon := \frac{e}{2mc^2}$$

with U_{acc} denoting the acceleration voltage of the source, C_c is the chromatic aberration of the objective lens, and α_c is the aperture angle of the beam furnished by the condenser.

Detection

Intensity Generated by a Single Electron

As the electron wave ψ reaches the image plane $\Gamma_{\text{imag}} = \boldsymbol{\omega}^\perp + p\boldsymbol{\omega}$ (Fig. 2), it forms an intensity distribution, a process that is encoded by the *intensity operator*:

$$\mathcal{I}(\psi)(\boldsymbol{x}) := \left|\psi(\boldsymbol{x} + p\boldsymbol{\omega})\right|^2 \quad \text{for } \boldsymbol{x} \in \boldsymbol{\omega}^\perp. \tag{26}$$

If the scattered part is a small perturbation $\triangle\psi$ to the incident wave ψ_0, then it is natural to *linearize the intensity*:

$$\mathcal{I}(\psi_0 + \triangle\psi)(\boldsymbol{x}) \approx \left|\psi_0(\boldsymbol{x} + p\boldsymbol{\omega})\right|^2 + \psi_0(\boldsymbol{x} + p\boldsymbol{\omega})\,\overline{\triangle\psi(\boldsymbol{x} + p\boldsymbol{\omega}}$$
$$+ \overline{\psi_0(\boldsymbol{x} + p\boldsymbol{\omega})}\,\triangle\psi(\boldsymbol{x} + p\boldsymbol{\omega}). \tag{27}$$

The Total Intensity and Its Detector Response

The micrograph is given by the intensity generated from *all* electrons reaching the image plane. In an ideal detection system, this would model the image in the micrograph, but all detection systems introduce noise and distortions which we now model.

First, one can treat each electron independently (independent electron assumption, see Sect. 4). Next, the models for the distribution of intensity and detector response are identical for each electron. Hence, in the absence of stochasticity, the micrograph is simply given by the detector response distribution for a single electron scaled by a factor corresponding to the number of electrons.

Detectors Types

Charged Coupled Device (CCD) cameras, or more precisely detectors based on slow-scan CCD sensors, have become the detection technique of choice in ET, so here we consider the modeling of such detectors.

Detection of electrons by a slow-scan CCD sensor can be divided into three separate stages [197]: (1) the conversion of incident electrons into photons in the scintillator (material that converts high-energy radiation like X-rays/electrons into visible light), (2) transport of photons from the scintillator to the CCD array (via fiber-optic or lens coupling), and (3) conversion of photons into electrons and the readout of the resulting digital signal. Each of these steps can be modeled at different levels of accuracy. The model outlined below is accurate enough for most computational treatments of micrographs.

Remark 11. It may seem strange to first convert an incident electron to photons, which are then converted back to an electric signal (electrons). The problem with using incident electrons directly to generate electron–hole pairs at the CCD is that they are too energetic, so they saturate the CCD, see, e.g., [146, p. 389]. In this context, it is worth mentioning the new generation detectors that directly detect the electron without an intermediate conversion to photons, and therefore posses significantly better sensitivity and noise properties [55].

Detector Response Distribution for a Single Electron

Consider a slow-scan CCD sensor placed at the image plane Γ_{imag}, so the scintillator entrance surface of the detector is Γ_{imag}. The *single electron intensity distribution*, which is the intensity distribution generated by a *single* electron at the scintillator entrance surface, is now given by

$$I(\boldsymbol{x}) := \mathcal{I}(\psi)(\boldsymbol{x}) \quad \text{for } \boldsymbol{x} \in \boldsymbol{\omega}^{\perp}. \tag{28}$$

The corresponding detector response distribution at the plane adjacent to the CCD, which models for the detector response per unit area (given as ADU per unit area) to a single electron in the absence of stochasticity, can then be modeled as [146, p. 391]

$$\boldsymbol{x} \mapsto C_{\text{gain}}(\boldsymbol{x})\Big\{I \underset{\boldsymbol{\omega}^{\perp}}{\circledast} \text{PSF}^{\text{det}}\Big\}(\boldsymbol{x}) + I_{\text{b}}(\boldsymbol{x}) \quad \text{for } \boldsymbol{x} \in \boldsymbol{\omega}^{\perp}. \tag{29}$$

Here, C_{gain} is the *overall gain* that measures the average number of digital counts that a single incident electron gives rise to, I_{b} is the distribution of the background signal, and PSF^{det} is the detector response function that models the spreading of the signal generated by a single incident electron within the scintillator and fiber-optic/lens coupling. All these components can be determined from specifications of the detector and/or estimated from specific calibration measurements (section on p. 982).

Total Detector Response Distribution

Let N_0 denote the *image dose*, which is defined as the average number of incident electrons per unit area at Γ_{imag} used when the micrograph is acquired. Then, (29) combined with the independent electron assumption allows us to model the total detector response distribution in the absence of stochasticity as

$$C(\boldsymbol{x}) := N_0 C_{\text{gain}}(\boldsymbol{x})\Big\{I \underset{\boldsymbol{\omega}^{\perp}}{\circledast} \text{PSF}^{\text{det}}\Big\}(\boldsymbol{x}) + N_0 I_{\text{b}}(\boldsymbol{x}) \quad \text{for } \boldsymbol{x} \in \boldsymbol{\omega}^{\perp}. \tag{30}$$

Characteristics of the Noise

To model the detector response for a pixel in the micrograph, we need to integrate the total detector response distribution in (30) over the pixel and model the stochasticity.

Shot Noise

The emission of electrons at the source (section "Illumination") is a stochastic process. This gives rise to *shot noise* whose influence is most severe in low-dose micrographs, such as in ET when one records multiple micrographs of biological cryo-specimens.

Shot noise is not Gaussian but depends on the signal level. It originates from the fact that emission of electrons is a Poisson process, so the actual image dose is a Poisson distributed random variable with mean equal to the nominal image dose N_0. To model shot noise, consider first the single electron intensity distribution I in (28). Each point $\boldsymbol{x} \in \boldsymbol{\omega}^{\perp}$ gives rise to an intensity distribution at the plane adjacent to the CCD:

$$\boldsymbol{y} \mapsto C_{\text{gain}}(\boldsymbol{x}) I(\boldsymbol{x}) \, \text{PSF}^{\text{det}}(\boldsymbol{x} - \boldsymbol{y}) + I_{\text{b}}(\boldsymbol{x}) \quad \text{for } \boldsymbol{y} \in \boldsymbol{\omega}^{\perp}.$$

Now, consider a micrograph acquired using a image dose of N_0 and let $\mathsf{C}(\triangle)$ denote the corresponding detector response from a (measurable) set $\triangle \subset \boldsymbol{\omega}^{\perp}$. Due to the Poisson stochasticity of the total dose, it is given by

$$\mathsf{C}(\triangle) = \int_{\triangle} \int_{\boldsymbol{\omega}^{\perp}} C_{\text{gain}}(\boldsymbol{x}) I(\boldsymbol{x}) \, \text{PSF}^{\text{det}}(\boldsymbol{x} - \boldsymbol{y}) + I_{\text{b}}(\boldsymbol{x}) \, \mathrm{d}\mathsf{N}(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{y}, \tag{31}$$

where N is a Poisson random measure (see [34, chapter VI]) with mean given by $N_0\,\mathrm{d}\boldsymbol{x}$, i.e.,

$$\mathsf{N}(\triangle) \sim \text{Poisson}\left[N_0 \int_{\triangle} 1\,\mathrm{d}\boldsymbol{x}\right]. \tag{32}$$

Remark 12. The expression in (31) models the noise in a continuum setting. The reason for insisting on a continuum model is that there is no natural pixelization of the image plane (scintillator entrance plane), which is where the Poisson stochasticity is realized since the electrons generate intensities in that plane. Similar models for Poisson distributed data when data sampling is in continuum are given in [83, 84].

Pixelization and Read-Out Noise

Besides the shot noise, there is additive and multiplicative noise. Both are related to a CCD pixel, so we next consider the pixelization in the CCD.

A pixelization in the CCD corresponds to a finite tessellation $\{\triangle_j\}_j$ of the bounded subregion of $\boldsymbol{\omega}^\perp$ that constitutes the CCD detector. The subsets $\triangle_j$ represent pixels and are typically given as square regions centered around some suitably chosen point $\boldsymbol{x}_j$ (often the midpoint). Disregarding multiplicative stochasticity, measured data from a single pixel can be modeled as a sample of the stochastic variable

$$\mathsf{C}^{\text{data}}(j) := \mathsf{C}(\triangle_j) + \mathsf{E}(j) \quad \text{for } j = 1, \ldots, n_{\text{det}}. \tag{33}$$

In the above, C is given by (31) and $\mathsf{E}(j) \sim \text{Normal}(\mu_j, \sigma_j)$ is a Gaussian random variable with mean μ_j and variance σ_j that models additive read-out noise, and n_{det} is the total number of pixels. The corresponding stochastic model for a micrograph is the stochastic n_{det}-vector

$$\boldsymbol{\mathsf{C}}^{\text{data}} := \left(\mathsf{C}^{\text{data}}(1), \ldots, \mathsf{C}^{\text{data}}(n_{\text{det}})\right). \tag{34}$$

Probability Distribution of Image Data

It is quite difficult to derive a closed form expression for the probability distribution of $\boldsymbol{\mathsf{C}}^{\text{data}}$ whose samples represent micrographs. One can however derive closed-form expressions for the expected value and covariance:

$$\begin{aligned}
\mathbf{E}\left[\mathsf{C}^{\text{data}}(j)\right] &= N_0 \int_{\triangle_j} \left[C_{\text{gain}}(\boldsymbol{x})\{I \underset{\boldsymbol{\omega}^\perp}{\circledast} \text{PSF}^{\text{det}}\}(\boldsymbol{x}) + N_0 I_{\text{b}}(\boldsymbol{x})\right] \mathrm{d}\boldsymbol{x} + \mu_j \\
\mathbf{Cov}\left[\mathsf{C}^{\text{data}}(j), \mathsf{C}^{\text{data}}(l)\right] &= N_0 \int_{\triangle_j} \int_{\triangle_l} \int_{\boldsymbol{\omega}^\perp} \Big[C_{\text{gain}}(\boldsymbol{x}) C_{\text{gain}}(\boldsymbol{y}) \\
&\qquad \text{PSF}^{\text{det}}(\boldsymbol{x} - \boldsymbol{y})\,\text{PSF}^{\text{det}}(\boldsymbol{z} - \boldsymbol{y}) I(\boldsymbol{y})\Big] \mathrm{d}\boldsymbol{y}\,\mathrm{d}\boldsymbol{z}\,\mathrm{d}\boldsymbol{x} + \mathbf{Cov}\left[\mathsf{E}(j), \mathsf{E}(l)\right].
\end{aligned} \tag{35}$$

The Measured Image Data

The *detector operator* $\mathcal{T}^{\text{det}}$ maps a single electron intensity distribution to the expected detector response distribution, so it is defined through the relation

$$\mathbf{E}\Big[\mathsf{C}^{\text{data}}(j)\Big] = \int_{\triangle_j} \mathcal{T}^{\text{det}}(I)(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} + \mu_j \tag{36}$$

where I is the aforementioned single electron intensity distribution. From (35) we get

$$\mathcal{T}^{\text{det}}(I)(\boldsymbol{x}) = N_0 C_{\text{gain}}(\boldsymbol{x})\big\{I \underset{\boldsymbol{\omega}^\perp}{\circledast} \text{PSF}^{\text{det}}\big\}(\boldsymbol{x}) + N_0 I_{\text{b}}(\boldsymbol{x}) \quad \text{for } \boldsymbol{x} \in \boldsymbol{\omega}^\perp, \tag{37}$$

so, $\mathcal{T}^{\text{det}}(I)$ equals (30), the total detector response distribution in the absence of stochasticity.

A common approximation is to replace the integration over $\triangle_j$ in (36) with a point-evaluation at a suitable point:

$$\mathbf{E}\Big[\mathsf{C}^{\text{data}}(j)\Big] \approx |\triangle_j|\big\{\mathcal{T}^{\text{det}}(I) \underset{\boldsymbol{\omega}^\perp}{\circledast} \text{PSF}_j^{\text{pix}}\big\}(\boldsymbol{x}_j) + \mu_j. \tag{38}$$

Here, $|\triangle_j|$ is the pixel area, $\boldsymbol{x}_j \in \triangle_j$ is some suitable evaluation point (typically the mid-point), and $\text{PSF}_j^{\text{pix}}$ is the associated pixel-shape function.

Remark 13. Further simplification of (38) is obtained when one considers measured data that has undergone basic pre-processing (section "Basic Pre-processing"). Then C_{gain} is constant and expected background signal μ_j can be estimated and deducted from data, so we may set $\mu_j = 0$. It is also common to consider the simplest pixel-shape function $\text{PSF}_j^{\text{pix}} = \delta_{\boldsymbol{x}_j}$. With these simplifications in place, the measured data associated with pixel j is modeled as

$$\mathbf{E}\Big[\mathsf{C}^{\text{data}}(j)\Big] \approx |\triangle_j| N_0 C_{\text{gain}}\big\{I \underset{\boldsymbol{\omega}^\perp}{\circledast} \text{PSF}^{\text{det}}\big\}(\boldsymbol{x}_j). \tag{39}$$

Forward Operator for Combined Phase and Amplitude Contrast

The forward operator is defined as $\mathcal{T}\colon \mathscr{X} \to \mathscr{H}$, where $\mathscr{X}$ is a vector space of functions $U\colon \Omega \to \mathbb{C}_+$ that represents a scattering potential (section on p. 952) and $\mathscr{H}$ is a vector space of functions $\boldsymbol{x} \mapsto g(\boldsymbol{\omega}, \boldsymbol{x}) \in \mathbb{R}_+$ for $\boldsymbol{x} \in \boldsymbol{\omega}^\perp$. Here, the function $g(\boldsymbol{\omega}, \cdot)$ represents the expected detector response distribution at the plane adjacent to the CCD (just prior to sampling and read-out) when the TEM image is acquired using incident electrons traveling along $\boldsymbol{\omega} \in S^2$.

In order to derive an expression for the forward operator, we first observe that by (6), (19), and (26), the single electron intensity distribution is given as

$$I(U)(\boldsymbol{\omega}, \boldsymbol{x}) = \mathcal{I} \circ \mathcal{T}^{\mathrm{op}} \circ \mathcal{T}^{\mathrm{sc}}_{\boldsymbol{\omega}}(U)(\boldsymbol{x}) \quad \boldsymbol{x} \in \boldsymbol{\omega}^{\perp}. \tag{40}$$

The forward operator corresponds to the expected detector response distribution, so by (36)–(37) we get

$$\begin{aligned} \mathcal{T}(U)(\boldsymbol{\omega}, \boldsymbol{x}) &:= \mathcal{T}^{\mathrm{det}}\Big(I(U)(\boldsymbol{\omega}, \cdot)\Big)(\boldsymbol{x}) \\ &\approx N_0(\boldsymbol{\omega}) C_{\mathrm{gain}} \Big\{ I(U)(\boldsymbol{\omega}, \cdot) \underset{\boldsymbol{\omega}^{\perp}}{\circledast} \mathrm{PSF}^{\mathrm{det}} \Big\}(\boldsymbol{x}) \quad \text{for } \boldsymbol{x} \in \boldsymbol{\omega}^{\perp}. \end{aligned} \tag{41}$$

Here, $N_0(\boldsymbol{\omega})$ denotes the image dose used to acquire the micrograph, I is the corresponding single electron intensity distribution given by (40), and the last approximate equality follows from setting $I_{\mathrm{b}} \equiv 0$ (Remark 13).

To get a more explicit expression for $\mathcal{T}$, we now consider various approximations of the single electron intensity distribution $I(U)$:

Lens-less imaging: This is the setting considered in section on p. 960 that results in the model (21) for $\mathcal{T}^{\mathrm{op}}$, so (40) becomes

$$I(U)(\boldsymbol{\omega}, \boldsymbol{x}) \approx \frac{k^2}{(2\pi)^4 (p-q)^2} \left| \Big\{ \mathrm{PSF}^{\mathrm{fsp}}_{p-q} \underset{\boldsymbol{\omega}^{\perp}}{\circledast} \mathcal{T}^{\mathrm{sc}}_{\boldsymbol{\omega}}(U)(q\boldsymbol{\omega} + \cdot) \Big\}(\boldsymbol{x}) \right|^2 \tag{42}$$

where $\mathrm{PSF}^{\mathrm{fsp}}_{p-q}$ is given by (22b).

Single thin lens with an aperture: This is the setting considered in section on p. 960 that results in the model (23) for $\mathcal{T}^{\mathrm{op}}$, so (40) becomes

$$I(U)(\boldsymbol{\omega}, \boldsymbol{x}) \approx \frac{1}{(2\pi)^4 M^2} \left| \Big\{ \mathrm{PSF}^{\mathrm{op}} \underset{\boldsymbol{\omega}^{\perp}}{\circledast} \mathcal{T}^{\mathrm{sc}}_{\boldsymbol{\omega}}(U)(q\boldsymbol{\omega} + \cdot) \Big\}\Big(\frac{\boldsymbol{x}}{M}\Big) \right|^2. \tag{43}$$

where $\mathrm{PSF}^{\mathrm{op}}$ is given by (24).

We conclude by providing the most common closed form expressions for $\mathcal{T}$ relevant for combined phase and amplitude contrast modeling in HRTEM imaging. These are based on combining specific models of the scattered wave $\mathcal{T}^{\mathrm{sc}}_{\boldsymbol{\omega}}(U)$ with the above expressions of the single electron intensity distribution.

Standard Phase Contrast Model

For TEM imaging of thin weakly scattering specimens, such as cryo-fixated biological specimens, one can combine a series of approximations that ultimately result in an affine forward model. Currently this is the best trade-off between computational feasibility and accuracy for ET on HRTEM data.

The first step is to employ the weak phase object approximation (17), so

$$\mathcal{T}^{\mathrm{sc}}_{\boldsymbol{\omega}}(U)(\boldsymbol{x} + q\boldsymbol{\omega}) \approx \psi_0(\boldsymbol{x} + q\boldsymbol{\omega}) \Big(1 + i\sigma \mathcal{P}(U)(\boldsymbol{\omega}, \boldsymbol{x})\Big) \quad \text{for } \boldsymbol{x} \in \boldsymbol{\omega}^{\perp}. \tag{44}$$

Next, assume the real and imaginary parts U_{re} and U_{im} of U are coupled like in (7). Insert (44) with (7) into (43), linearize the intensity as in (27), and use the fact that ψ_0 is a plane wave, so $|\psi_0|^2 \equiv 1$. Then, as worked out in [51, eqs. (37)–(38)], the single electron intensity distribution (43) becomes

$$I(U)(\boldsymbol{\omega}, \boldsymbol{x}) \approx \frac{1}{M^2}\left(1 - \frac{2\sigma}{(2\pi)^2}\left\{\mathrm{PSF}^{\mathrm{op}}_{\mathrm{tot}}(\boldsymbol{\omega}, \cdot) \underset{\boldsymbol{\omega}^\perp}{\circledast} \mathcal{P}(U_{\mathrm{re}})(\boldsymbol{\omega}, \cdot)\right\}\left(\frac{\boldsymbol{x}}{M}\right)\right). \tag{45}$$

Here, $\mathrm{PSF}^{\mathrm{op}}_{\mathrm{tot}}(\boldsymbol{\omega}, \boldsymbol{z}) := \mathrm{PSF}^{\mathrm{op}}_{\mathrm{im}} + Q(\boldsymbol{\omega})\,\mathrm{PSF}^{\mathrm{op}}_{\mathrm{re}}$, with $\mathrm{PSF}^{\mathrm{op}}_{\mathrm{re}}$ and $\mathrm{PSF}^{\mathrm{op}}_{\mathrm{im}}$ denoting the real and imaginary parts of $\mathrm{PSF}^{\mathrm{op}}$ in (24), i.e.,

$$\mathrm{PSF}^{\mathrm{op}}_{\mathrm{im}}(\boldsymbol{x}) = \mathcal{F}_{\boldsymbol{\omega}^\perp}\left[A_\Sigma\left(\frac{f}{k}\cdot + f\boldsymbol{\omega}\right)\sin\left[W\left(|\cdot|^2\right)\right]\right](\boldsymbol{x})$$

$$\mathrm{PSF}^{\mathrm{op}}_{\mathrm{re}}(\boldsymbol{x}) = \mathcal{F}_{\boldsymbol{\omega}^\perp}\left[A_\Sigma\left(\frac{f}{k}\cdot + f\boldsymbol{\omega}\right)\cos\left[W\left(|\cdot|^2\right)\right]\right](\boldsymbol{x}).$$

Finally, inserting (45) into (41) gives the following expression for the forward operator:

$$\mathcal{T}(U)(\boldsymbol{\omega}, \boldsymbol{x}) \approx \frac{N_0(\boldsymbol{\omega})C_{\mathrm{gain}}}{M^2}\left[\left\{\mathrm{PSF}^{\mathrm{det}} \underset{\boldsymbol{\omega}^\perp}{\circledast} 1\right\}(\boldsymbol{x}) - \frac{2\sigma}{(2\pi)^2}\left\{\mathrm{PSF}^{\mathrm{tot}}(\boldsymbol{\omega}, \cdot) \underset{\boldsymbol{\omega}^\perp}{\circledast} \mathcal{P}(U_{\mathrm{re}})(\boldsymbol{\omega}, \cdot)\right\}\left(\frac{\boldsymbol{x}}{M}\right)\right] \tag{46}$$

where $\mathrm{PSF}^{\mathrm{tot}}(\boldsymbol{\omega}, \boldsymbol{x}) := \left\{\mathrm{PSF}^{\mathrm{det}} \underset{\boldsymbol{\omega}^\perp}{\circledast} \left(\mathrm{PSF}^{\mathrm{op}}_{\mathrm{im}} + Q(\boldsymbol{\omega})\,\mathrm{PSF}^{\mathrm{op}}_{\mathrm{re}}\right)\right\}(\boldsymbol{x})$ for $\boldsymbol{x} \in \boldsymbol{\omega}^\perp$.

Phase Contrast Model with Lens-Less Imaging

One can also consider the simpler lens-less imaging model (42) for the intensity. Combine the weak phase object approximation (44) with (7), and insert that into (42), linearize the intensity as in (27), and use the fact that ψ_0 is a plane wave, so $|\psi_0|^2 \equiv 1$. The end result is an expression of the type (46) but with a different PSF.

Phase Contrast Model with Ideal Detector Response and Optics

This is the case when $A_\Sigma \equiv 1$, $W \equiv 0$, and $\mathrm{PSF}^{\mathrm{det}} \equiv \delta_{\boldsymbol{\omega}^\perp}$ in (46). Then, $\mathrm{PSF}^{\mathrm{op}}_{\mathrm{im}} \equiv 0$ and $\mathrm{PSF}^{\mathrm{op}}_{\mathrm{re}} = \mathrm{PSF}^{\mathrm{det}} \equiv \delta_{\boldsymbol{\omega}^\perp}$, so

$$\mathcal{T}(U)(\boldsymbol{\omega}, \boldsymbol{x}) \approx \frac{N_0(\boldsymbol{\omega})C_{\mathrm{gain}}}{M^2}\left(1 - \frac{2\sigma}{(2\pi)^2}Q(\boldsymbol{\omega})\mathcal{P}(U_{\mathrm{re}})\left(\boldsymbol{\omega}, \frac{\boldsymbol{x}}{M}\right)\right). \tag{47}$$

Remark 14. If $Q(\boldsymbol{\omega}) = 0$, then there is no detectable signal from the specimen. This is to be expected since the optics is needed to image phase contrast.

Forward Operator for Amplitude Contrast Only

Here we ignore the phase shift imposed by the specimen, so all the contrast in the micrograph is assumed to arise from electrons that are blocked from reaching the image plane (amplitude contrast). This is adequate for modeling essentially all medium-resolution contrast (beyond 2–3 nm) seen in micrographs, e.g., contrast from negatively stained specimens. Furthermore, all the early, successful attempts at 3D reconstruction relied on such a model [43, 97].

The Intensity

The starting point for modeling amplitude contrast is to consider electrons as particles and use a model similar to the Lambert–Beer law for the attenuation of X-rays in computed tomography. Thus, consider an incident beam of electrons traveling along $\boldsymbol{\omega} \in S^2$ and let $\mu_\alpha : \mathbb{R}^3 \to \mathbb{R}_+$ denote the mass attenuation coefficient of the specimen for an angle α. Hence, $\mu_\alpha(\boldsymbol{x})$ is proportional to the probability that an incident electron traveling along $\boldsymbol{\omega}$ scatters with an angle larger than α at $\boldsymbol{x}$. A large value therefore implies that the electrons scatter with high angles and a small value means that the specimen is relatively transparent to the electron beam.

After interacting with the specimen, the incident beam gives rise to a (total) intensity distribution at the scintillator entrance surface $\boldsymbol{x} \mapsto N_0 I(\boldsymbol{x})$ where N_0 is the incoming dose and I is the single electron intensity distribution defined in (28). The latter is now modeled as [176, p. 157]:

$$I(\boldsymbol{x}) = I_0(\boldsymbol{x}) \exp\big(-\mathcal{P}(\mu_\alpha)(\boldsymbol{\omega}, \boldsymbol{x})\big) \quad \text{for } \boldsymbol{x} \in \boldsymbol{\omega}^\perp. \tag{48}$$

Here, $\mathcal{P}$ is the ray transform in (1) and I_0 is the single electron intensity distribution one would obtain in the absence of a specimen.

The Forward Operator

The forward operator is obtained in the same way as (41) is obtained from (40), i.e.,

$$\mathcal{T}(\mu_\alpha)(\boldsymbol{\omega}, \boldsymbol{x}) \approx N_0(\boldsymbol{\omega}) C_{\text{gain}} \Big\{ I(\mu_\alpha)(\boldsymbol{\omega}, \cdot) \underset{\boldsymbol{\omega}^\perp}{\circledast} \text{PSF}^{\text{det}} \Big\}(\boldsymbol{x}) \text{ for } \boldsymbol{x} \in \boldsymbol{\omega}^\perp. \tag{49}$$

Here, $I(\mu_\alpha)(\boldsymbol{\omega}, \cdot)$ is given by the left-hand side of (48), $N_0(\boldsymbol{\omega})$ is the image dose, and $\mu_\alpha : \mathbb{R}^3 \to \mathbb{R}_+$ fully characterizes the specimen for scattering with angles up to α. Hence, if (49) holds, one can pre-process data so that the corresponding forward operator equals the ray transform $\mathcal{P}$ in (1).

Connection to Quantum Mechanics

The particle model in (48) can be related to the wave model based on the Schrödinger equation (5). To do this we need to take a closer look at μ_α.

First, in the amplitude contrast only model the phase contrast is ignored, so ψ_{sc} and ψ_0 have the same phase. Since $I = |\psi_{\text{sc}}|^2$, (48) yields

$$\psi_{\text{sc}}(\boldsymbol{x}) = \psi_0(\boldsymbol{x}) \exp\Big(-\frac{1}{2} \int_{-\infty}^{\infty} \mu_\alpha(\boldsymbol{x} + s\boldsymbol{\omega})\, \mathrm{d}s \Big).$$

Next, consider the projection approximation (16) for the ψ_{sc}. Equating the above expression for ψ_{sc} with the one in (16) yields $\mu_\alpha = 2\sigma U_{\text{im}}$ with $\sigma := me/k\hbar^2$. This relates the linear attenuation coefficient μ_α to the imaginary part U_{im} of the potential U in (2). This relation however only makes sense if one removes the explicit model for the aperture in the phase contrast model (simply set $A_\Sigma \equiv 1$ in (24)). Then, U_{im} acts as an attenuation term that accounts for both aperture and inelastic scattering.

Another connection is provided through the (elastic and inelastic) scattering cross-section. From [176, eq. (6.1)] we get that

$$\mu_\alpha(\boldsymbol{x}) = \frac{\rho(\boldsymbol{x}) N_A}{M(\boldsymbol{x})} \sigma_\alpha\Big(\frac{\boldsymbol{x}}{|\boldsymbol{x}|}\Big) \quad \text{for } \boldsymbol{x} \in \mathbb{R}^3,$$

where $\sigma_\alpha \colon S^2 \to \mathbb{R}_+$ represents the local (elastic and inelastic) scattering cross-section, $\rho \colon \mathbb{R}^3 \to \mathbb{R}_+$ is the density, $M \colon \mathbb{R}^3 \to \mathbb{R}_+$ is the local molecular weight (in mol per unit volume), and N_A is Avogadro's constant (in mol per unit mass). Next, by [36, theorem 2.6], we can express the solution to (4a)–(4c) with $\boldsymbol{A} \equiv \boldsymbol{0}$ as

$$\psi(\boldsymbol{x}) \approx \exp(ik\boldsymbol{x} \cdot \boldsymbol{\omega}) + \psi_\infty\Big(\frac{\boldsymbol{x}}{|\boldsymbol{x}|}\Big) \frac{\exp(ik|\boldsymbol{x}|)}{|\boldsymbol{x}|}.$$

The term $\psi_\infty \colon S^2 \to \mathbb{C}$ is called the far-field pattern and $i\psi_\infty$ is the complex scattering amplitude, which in turn is related to the local scattering cross section by [78, eq. (6)]

$$\sigma_\alpha(\boldsymbol{x}) = \int_{S^2 \setminus \Sigma_\alpha} \Big| \psi_\infty\Big(\frac{\boldsymbol{x}}{|\boldsymbol{x}|}\Big) \Big|^2 \, \mathrm{d}\sigma(\boldsymbol{x}).$$

In the above, $\Sigma_\alpha \subset S^2$ is the region defined by the aperture (i.e., the angular range that is blocked by the aperture).

Summary

The model for the intensity depends on the type of contrast one seeks to model. The model presented in section "Forward Operator for Amplitude Contrast Only" for the amplitude contrast is only valid for thin specimens, e.g., unstained biological

specimens imaged using 100 keV electrons have to be thinner than 10 nm. If the specimen is thicker, this approach is only applicable to model low resolution image features.

At higher resolution, like in HRTEM imaging, one needs to account for the phase contrast, leading to the combined phase and amplitude contrast model (section "Forward Operator for Combined Phase and Amplitude Contrast"). This modeling framework depends highly on the model for the scattering operator $\mathcal{T}_{\omega}^{\mathrm{sc}}$ defined in (6). In section "Electron–Specimen Interaction" we have presented several approximations to the scattering operator. The most accurate model is the one given by the multi-slice method that accounts for multiple scattering and does not assume weakly scattering specimen. The downside is that this results in a nonlinear model for image formation that is computationally challenging within an inverse problems setting. The first order Born approximation leads to an affine model that is more accurate than the geometrical optics approximation. It is valid whenever $kd\triangle U \ll \pi$ where d is the size of features one seeks to resolve and $\triangle U$ is the contrast (actually it is the ray transform of the contrast) [135]. Its implementation is however more complex requiring usage of nonuniform fast Fourier transform techniques for efficient evaluation of the scattering operator. Finally, for approaches based on geometrical optics approximation (13), only the weak phase object approximations (17) result in an affine model.

Currently there is no computationally feasible theory for modeling 2D image features in the micrographs at an intermediate resolution range (<1 nm) in an amorphous specimen that is not ultra thin (thickness between 15 and 20 nm) [176, p. 157]. Models for simulation of such micrographs, see, e.g., [176, section 6.7], are computationally to demanding to be used within an 3D image reconstruction scheme. This is also confirmed by a recent study in [189] that considers cryo-electron microscopy (section on p. 942) of amorphous biological specimens. There, one concludes that the projection and weak-phase object approximations are satisfied for simulation of TEM image features not smaller than 0.5 nm, so higher resolution simulations require more advanced models for the scattering operator, see also [78]. In ET life-science applications one typically does not attempt to interpret 3D image features that are smaller than 3 nm, so a model for the scattering operator that is to be used within a tomographic reconstruction scheme can be based on the weak-phase object approximation.

5 Data Acquisition Geometry

The reconstruction problem in ET is an example of a tomographic inverse problem. Data naturally divides into sub-datasets, each containing data acquired when the object has undergone a specific Euclidean transformation (tomographic data). The associated *data acquisition geometry* is a specification of the Euclidean transformations that describe how the object moves in between the acquisition of the sub-datasets.

In ET each sub-dataset corresponds to a micrograph, a 2D TEM image, acquired using uniform illumination mode (Sect. 2). A *tilt-series* refers to the series of micrographs, and the *zero-tilt* micrograph is the one recorded when the specimen input plane is as orthogonal as possible to the optical axis. The specimen is only rotated (no translations) in-between sub-datasets, so the data acquisition geometry follows a *parallel beam geometry*, i.e., all electrons that generate an image travel along the same direction and the detector is planar and orthogonal to this incident direction of propagation.

Parallel Beam Geometries

For parallel beam geometry, continuum tomographic data is represented as a real-valued function defined on the tangent bundle $T(S^2)$:

$$T(S^2) := \{(\boldsymbol{\omega}, \boldsymbol{x}) \in S^2 \times \mathbb{R}^3 : \boldsymbol{x} \in \boldsymbol{\omega}^{\perp}\}.$$

$T(S^2)$ provides natural coordinates for tomographic data since $(\boldsymbol{\omega}, \boldsymbol{x}) \in T(S^2)$ represents data at $\boldsymbol{x}$ in the detector plane when the orientation of the detector is given by $\boldsymbol{\omega} \in S^2$ (normal to the detector plane).

In ET data is given on a 3D subset of $T(S^2)$, typically be restricting $\boldsymbol{\omega} \in S^2$ to a curve $S_0 \subset S^2$. Hence, the subsets of $T(S^2)$ we consider are of the following form:

Definition 1. *A parallel beam line complex* (also called Orlov's pencil) is a subset $\mathcal{M}_{S_0} \subset T(S^2)$ where

$$\mathcal{M}_{S_0} := \big\{(\boldsymbol{\omega}, \boldsymbol{x}) \in T(S^2) : \boldsymbol{\omega} \in S_0\big\} \quad \text{for fixed } S_0 \subset S^2.$$

S_0 is sometimes called the *angular data collection aperture* [42]. Next, for a given region of interest $\Omega_0 \subset \mathbb{R}^3$, we define

$$\mathcal{M}_{S_0}(\Omega_0) := \big\{(\boldsymbol{\omega}, \boldsymbol{x}) \in \mathcal{M}_{S_0} : \boldsymbol{x} = \boldsymbol{y} - (\boldsymbol{y} \cdot \boldsymbol{\omega})\boldsymbol{\omega} \text{ for some } \boldsymbol{y} \in \Omega_0.\big\}.$$

For parallel beam geometries, continuum tomographic data is represented by a function $g : \mathcal{M}_{S_0} \to \mathbb{R}$ and the *data acquisition geometry* is a specification of the curve S_0.

Examples Relevant for ET

Here we give some examples of parallel beam geometries that occur in ET. The corresponding curves S_0 are shown in Fig. 3. For single- and double-axis tilting, coordinates (x, y, z) in $\mathbb{R}^3$ are chosen so that the x-axis is parallel to the tilt-axis and the z-axis parallel to the electron beam when the specimen is in the zero-tilt position.

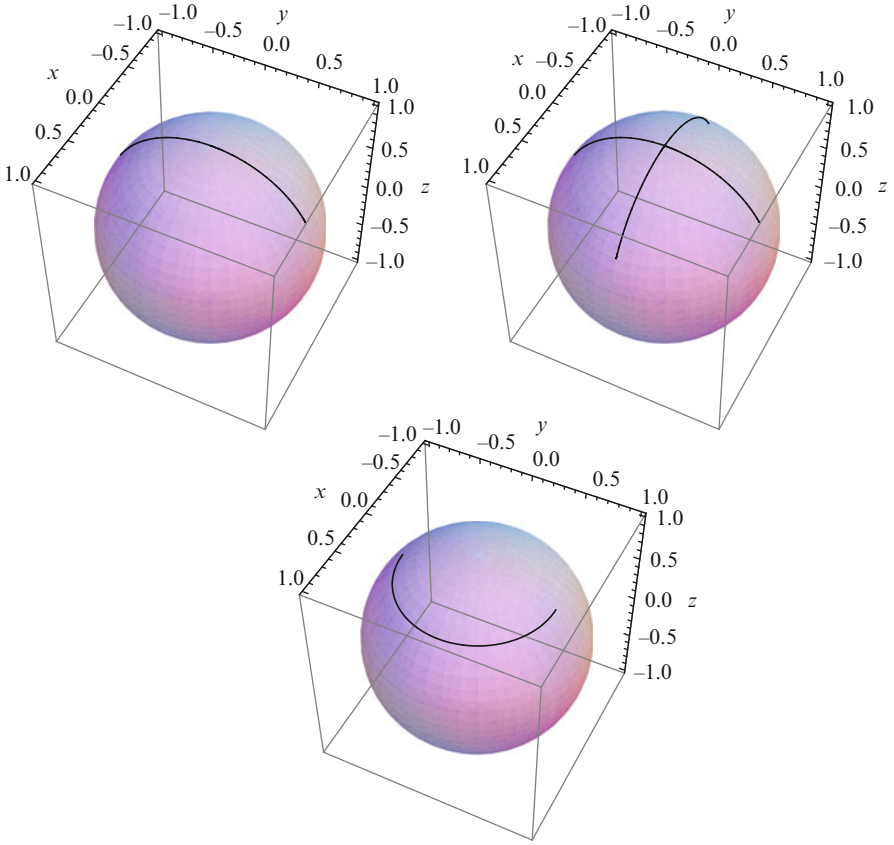


Fig. 3 Illustrating $S_0 \subset S^2$ for the three most common data collection geometries in ET, single-axis titling (*left*), double-axis tilting (*middle*), and conical-axis tilting (*right*). The z-axis is the electron beam direction, and in the case of single-tilting, the specimen is tilted around the x-axis

Single-axis tilting: Here $S_0 := \{(0, \sin\theta, \cos\theta) : -\theta_{\max} < \theta < \theta_{\max}\}$ where a typical value of the tilt-angle is $\theta_{\max} = 60°$.

Double-axis tilting: S_0 is here the union of two single-axis tilting geometries, one along the x- and the other along the y-axis. As for single-axis tilting, a typical value for the tilt-angle is $\theta_{\max} = 60°$.

Conical-axis tilting: Fix β, typically $\beta = 45°$, and define

$$S_0 := \{(\sin\beta \sin\theta, \sin\beta \cos\theta, \cos\theta) : 0 < \theta < 180°\}.$$

In the ET community, the curve S_0 is often parametrized in terms of the Euler angles [143, p. 6] instead of describing it, like above, as a curve on S^2. This is also the convention used by most software packages for ET.

6 The Reconstruction Problem in ET

Mathematical Formulation

We will state two variants of the 3D reconstruction problem in ET, one for continuum noise-free tomographic data and one for finitely sampled noisy tomographic data.

The Specimen

In the combined phase and amplitude contrast model (section "Forward Operator for Combined Phase and Amplitude Contrast"), the specimen is fully characterized by its 3D scattering potential $U_{\text{true}}: \mathbb{R}^3 \to \mathbb{C}_+$. In the amplitude contrast model (section "Forward Operator for Amplitude Contrast Only"), it is fully characterized by its 3D mass attenuation coefficient $\mu_\alpha: \mathbb{R}^3 \to \mathbb{R}_+$. Since the mass attenuation coefficient can be identified with the imaginary part of the scattering potential (paragraph on p. 970), we will henceforth use the same notation, U_{true}, for both these functions and refer to them as the *signal*.

Next, the specimen is contained in an infinite slab $\Omega \subset \mathbb{R}^3$ given as an unbounded region enclosed between two parallel hyperplanes, Γ_{in} (specimen input plane) and Γ_{out} (specimen exit plane). We also introduce the region of interest $\Omega_0 \subset \Omega$, which is a bounded domain where we seek to recover the true signal U_{true}. Note that U_{true} has compact support in Ω, but its support is not necessarily contained in Ω_0. In ET this is not the case (section "Incomplete Data, Uniqueness and Stability"). Finally, we introduce $\mathscr{X}$ (reconstruction space), the vector space of feasible signals. Elements in $\mathscr{X}$ are complex-valued functions with compact support in Ω with reasonable regularity (section on p. 952).

Continuum Tomographic Data

Here tomographic data is represented by a function $g: \mathcal{M}_{S_0}(\Omega_0) \to \mathbb{R}_+$ where $\mathcal{M}_{S_0}(\Omega_0) \subset T(S^2)$ given as in Definition 1. The data space $\mathscr{H}$ is a suitable set of such functions, typically one requires that it is contained in $\mathscr{S}\big(\mathcal{M}_{S_0}(\Omega_0), \mathbb{R}\big)$.

The inverse problem in ET is to recover the signal $U_{\text{true}} \in \mathscr{X}$ on Ω_0 from continuum tomographic data $g: \mathcal{M}_{S_0}(\Omega_0) \to \mathbb{R}_+$ where

$$g(\boldsymbol{\omega}, \boldsymbol{x}) = \mathcal{T}(U_{\text{true}})(\boldsymbol{\omega}, \boldsymbol{x}) \quad \text{for } (\boldsymbol{\omega}, \boldsymbol{x}) \in \mathcal{M}_{S_0}(\Omega_0). \tag{50}$$

Here $\mathcal{T}: \mathscr{X} \to \mathscr{H}$ is the forward operator associated with TEM imaging (section "Forward Operator for Combined Phase and Amplitude Contrast" for combined phase and amplitude contrast and section "Forward Operator for Amplitude Contrast Only" for amplitude contrast only). Finally, ET is local tomography problem, meaning that the support of U_{true} is not contained in Ω_0.

Sampled Tomographic Data

Finitely sampled tomographic data directly corresponds to a tilt-series. To each micrograph in the tilt-series, we associate a directional vector $\boldsymbol{\omega}_i \in S_0$ an image dose $N_0(\boldsymbol{\omega}_i)$. The former is given by the angle between the incident beam and the specimen input plane Γ_{in} (zero-tilt is when $\boldsymbol{\omega}_i \perp \Gamma_{\text{in}}$), and the latter is the total dose used to acquire the micrograph. Furthermore, there is a pixelization $\{\triangle_j\}_j$ associated with the CCD chip. If there are m_{tlt} micrographs in the tilt-series and the CCD chip has n_{det} pixels, then tomographic data is represented as a vector in $\mathbb{R}_+^m$ with $m = m_{\text{tlt}} n_{\text{det}}$.

Next, we introduce the sampled forward operator $\mathcal{T}\colon \mathscr{X} \to \mathbb{R}^m$ as

$$\mathcal{T}(U)(i,j) := \mathbf{E}\left[\mathsf{C}^{\text{data}}(i,j)\right] \quad \text{for } i = 1,\ldots,m_{\text{tlt}} \text{ and } j = 1,\ldots,n_{\text{det}} \tag{51}$$

where $\mathsf{C}^{\text{data}}(i,j)$ is the stochastic model in (33) for the measured value at the j:th pixel in the i:th micrograph (the dependency on U is suppressed in the notation). From (36) we can express the sampled forward operator in terms of the forward operator $\mathcal{T}\colon \mathscr{X} \to \mathscr{H}$ (section "Forward Operator for Combined Phase and Amplitude Contrast" for combined phase and amplitude contrast and section "Forward Operator for Amplitude Contrast Only" for amplitude contrast only):

$$\begin{aligned}\mathcal{T}(U)(i,j) &= \int_{\triangle_j} \mathcal{T}(U)(\boldsymbol{\omega}_i, \boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \approx |\triangle_j| \left\{\mathcal{T}(U)(\boldsymbol{\omega},\cdot) \underset{\boldsymbol{\omega}^\perp}{\circledast} \mathrm{PSF}_j^{\text{pix}}\right\}(\boldsymbol{x}_j) + \mu_j \\ &\approx |\triangle_j| \mathcal{T}(U)(\boldsymbol{\omega}, \boldsymbol{x}_j).\end{aligned} \tag{52}$$

The last two approximations follow from (38) and (39), respectively.

The corresponding inverse problem is to recover the signal $U_{\text{true}} \in \mathscr{X}$ on Ω_0 from a tilt-series $\boldsymbol{g} \in \mathbb{R}_+^m$ where

$$\boldsymbol{g} = \mathcal{T}(U_{\text{true}}) + \boldsymbol{g}^{\text{noise}}. \tag{53}$$

In the above, $\mathcal{T}$ is the sampled forward operator and $\boldsymbol{g}^{\text{noise}}(i,j)$ (noise component of data) is a sample of the random variable $\mathsf{C}^{\text{data}}(i,j) - \mathcal{T}(U_{\text{true}})(i,j)$.

Fully Discrete Setting

This is the case when the reconstruction space $\mathscr{X}$ is discretized by mapping an element in $\mathscr{X}$ to a vector in $\mathbb{R}_+^n$. The inverse problem is to recover the n-vector $\boldsymbol{U}_{\text{true}} \in \mathbb{R}^n$ from sampled tomographic data $\boldsymbol{g} \in \mathbb{R}_+^m$ where

$$\boldsymbol{g} = \boldsymbol{T}(\boldsymbol{U}_{\text{true}}) + \boldsymbol{g}^{\text{noise}}. \tag{54}$$

In the above, $\boldsymbol{T}\colon \mathbb{R}^n \to \mathbb{R}^m$ is the fully discretized forward operator that corresponds to $\mathcal{T}$.

Notion of Solution

An issue that often arise when dealing with inverse problems with noisy data and/or uncertain forward operator, like in ET, is that data is not in the range of the forward operator (inconsistent data). For such cases, (53) (or its fully discretized version (54)) will not have a solution (non-existence). A common approach to address this issue is to introduce a relaxed notion of solution by considering

$$\operatorname*{argmin}_{U \in \mathscr{X}} \mathcal{D}\Big(\boldsymbol{\mathcal{T}}(U), \boldsymbol{g}\Big) \quad \text{for given} \quad \mathcal{D}\colon \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}_+. \tag{55}$$

In the above, $\mathcal{D}$ is the *data discrepancy* that quantifies the goodness-of-fit. Choosing it as the 2-norm in $\mathbb{R}^m$ yields least-squares solutions, and more generally, deriving it from the data likelihood yields Maximum-Likelihood (ML) solutions. Existence for (53) with the above notion of a solution translates into existence of solutions to (55), see [196] for an overview of such results. In general, such results are highly dependent on the (i) space $\mathscr{X}$, (ii) data discrepancy $\mathcal{D}$, and (iii) forward operator $\boldsymbol{\mathcal{T}}\colon \mathscr{X} \to \mathbb{R}^m$.

Besides existence, one also has the issue of uniqueness. As with most inverse problems involving function spaces and finite data, if the inverse problem (53) has a solution, it will not be unique. Furthermore, in ET the number of data points is almost always less than the number of 3D voxels, so the same also holds for the fully discretized version (54). Hence, assuming (55) has solutions, there are infinitely many of them (non-uniqueness). Handling the issue of non-uniqueness is the art of regularization, where the basic idea is to make use of a priori knowledge about $U_{\text{true}} \in \mathscr{X}$.

Expressions for the Data Discrepancy

Choosing $\mathcal{D}$ as a suitable affine transformation of the negative log-likelihood of the random variable in (51) (that models measured data $\boldsymbol{g}$) allows one to interpret (55) as Maximum-Likelihood (ML) estimation of (53).

Retaining the above interpretation of (55), data with additive Gaussian noise (with zero mean and known covariance matrix $\boldsymbol{\Sigma}$) leads to choosing $\mathcal{D}$ as the Mahalanobis distance. It weights the usual 2-norm with the covariance matrix $\boldsymbol{\Sigma}$ of data to make the distance metric scale-invariant:

$$\mathcal{D}\Big(\boldsymbol{\mathcal{T}}(U), \boldsymbol{g}\Big) := \frac{1}{2}\big(\boldsymbol{\mathcal{T}}(U) - \boldsymbol{g}\big)^t \cdot \boldsymbol{\Sigma}^{-1} \cdot \big(\boldsymbol{\mathcal{T}}(U) - \boldsymbol{g}\big).$$

On the other hand, for Poisson data, $\mathcal{D}$ is given by the Kullback–Leibler divergence in the data space:

$$\mathcal{D}\Big(\mathcal{T}(U), \boldsymbol{g}\Big) := \sum_{i=1}^{m} \Big(g_i \log \mathcal{T}(U)_i - \mathcal{T}(U)_i\Big).$$

Now, the ET noise model in (34) is more complex. Deconvolving the detector noise leads to data that is Poisson with additive Gaussian noise. It turns out that the corresponding log-likelihood has a closed form expression [16]. If one insists on a closed form expression for the data discrepancy $\mathcal{D}$, this is probably the most accurate noise model for images acquired using a low photon/electron count and high detector noise. A difficulty however is that $\mathcal{D}$ becomes non-convex and non-smooth. Therefore, a variety of approximations have been suggested, like [15, 32] and [67, 110], the latter two based on formulating a Mahalanobis distance that approximates the log-likelihood.

7 Specific Difficulties in Addressing the Inverse Problem

The inverse problem in ET is associated with a number of difficulties that make it a challenging problem.

The Dose Problem

This is *the single most important* factor limiting the usefulness of ET in life sciences. As explained below, the dose problem arises due to damage induced by the electron–specimen interaction and limits the total number of micrographs that can be recorded.

When a specimen is irradiated by an electron beam, it gets progressively damaged due to ionization. The actual mechanisms underlying specimen damage are more complex and outside the scope of this review, the reader may consult [9, 46] and the references therein for further information. For our purposes, the implication of specimen damage is that the number of electrons used to irradiate the specimen need to be low enough to preserve the structural integrity of the specimen. For biological materials, the total dose varies between 1,000 and 10,000 e^-/nm^2 [148, section 3.2], depending on the size of image features that are of interest, the type of specimen, and the sample preparation method. Similar thresholds are presented by [103, section 3.2] and [19]. Hence, we are roughly dealing with a *total* of 300–3,000 e^-/pixel (at 25,000× magnification using a detector with a pixel-size of 14 μm) distributed over 60 or 120 images, so micrographs have very low signal-to-noise ratio with significant influence of shot noise, see also Fig. 4.

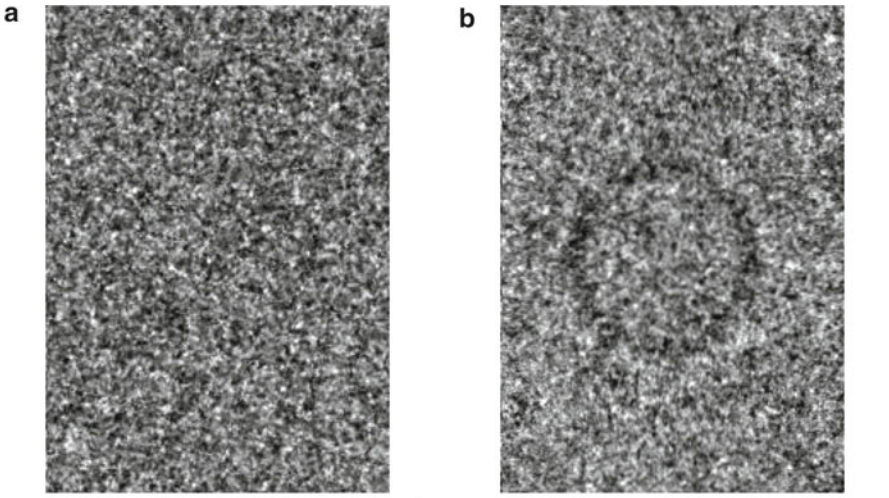

a
b
Low-dose (about 3000 e⁻/nm²) micrograph recorded at 1 μm under-focus.
High-dose micrograph taken at 12 μm under-focus at the end of the ET data collection.

Fig. 4 Micrographs of an in vitro specimen containing microsomes (small spherical shells with diameter $\sim$ 60 nm) recorded using a 200 kV TEM. Both micrographs represent the same rectangular subregion (H×W: 150×106 nm) of a zero-tilt micrograph. Note that the microsome, which is visible in the high-dose micrograph (**b**), is not visible in the low-dose micrograph (**a**). The high-dose micrograph is not used for reconstruction due to specimen damage. (**a**) Low-dose (about 3,000 e^-/nm^2) micrograph recorded at 1 μm under-focus. (**b**) High-dose micrograph taken at 12 μm under-focus at the end of the ET data collection

As a final note, the dose problem is less of an issue for ET applied to material sciences since those specimens are not as sensitive to damage induced by the electron–specimen interaction as biological specimens.

Incomplete Data, Uniqueness, and Stability

Mathematical uniqueness and stability results are only meaningful for inverse problems with continuum data (50) or in the fully discrete setting (54). For both cases, results are highly dependent on the forward operator, the data collection geometry, and the choice of the reconstruction and data spaces. Here we focus on (50), the case with continuum data.

It is easy to see that the inverse problem (50) cannot have a unique solution. One source for non-uniqueness is the phase problem (section "Phase Retrieval"), another is the fact that tomographic data is local, e.g., the support of U_{true} is not contained in Ω_0 whereas data is given only on $\mathcal{M}_{S_0}(\Omega_0)$. Furthermore, even if one can address the issue of non-uniqueness, there are strong indications that the inverse problem is severely ill-posed due to limitations in the data acquisition geometry (limited angle

data). These claims are supported by first considering the situation in a simplified setting, namely the standard phase contrast and the amplitude contrast models.

Standard Phase Contrast Model

The forward operator is given by (46), i.e., a ray transform followed by convolutions. Since the real and imaginary parts of the potential are related as in (7), any non-uniqueness issues associated with the phase problem are resolved. Hence, uniqueness rests upon the ability to uniquely invert the ray transform on $\mathcal{M}_{S_0}(\Omega_0)$, and stability refers to the aforementioned inversion and the deconvolution of the optics and detector PSFs.

Uniqueness

If the support of U_{true} is contained in the region of interest Ω_0, then uniqueness for inversion of the ray transform follows from [122, theorem 3.144] whenever S_0 is infinite (uncountable), no matter how small it is. On the other hand, there is non-uniqueness for any finite set of directions, no matter how many [122, corollary 3.147]. As already mentioned on p. 974, ET is a local tomography problem. Hence, without additional prior information about U_{true}, the ray transform is not uniquely invertible [134, section VI.4]. One option is then to reconstruct features that are uniquely recoverable (section on p. 994).

Stability

Even when U_{true} is supported in Ω_0 (so there is uniqueness), we still have the issue of stability. Uniqueness is by analytic continuation, a non-constructive method that is highly vulnerable to noise in the data. Hence, this approach cannot be used to give reliable information about U_{true} from data.

Now, any reasonable form of stability is based on the possibility to obtain stability estimates in Sobolev spaces, analogous to those for the complete data case $S_0 = S^2$, which in turn requires that $S_0 \subset S^2$ fulfills Orlov's criteria (every great circle on S^2 has a nonzero intersection with S_0) [42, section 2.3] [134, chapter VI] [141, chapter 6]. This does not hold for the parallel beam data collection geometries that occur in ET (section "Examples Relevant for ET"), leading to severe ill-posedness. The ill-posedness is easy to see in the single-axis tilting scheme, where the inversion of the ray transform reduces to inversion of the 2D Radon transform in planes orthogonal to the tilt-axis. The limitation on the tilt-angle means that each such 2D Radon inversion problem is a limited angle problem, which is known to be severely ill-posed [40, 114]. Finally, the deconvolution in ET of the optics and detector PSFs are unstable operations, further adding to the ill-posedness of the inverse problem (50).

Amplitude Contrast Model

The forward operator is (49), and the deconvolution of the detector response PSF^{det} does not affect uniqueness. Hence, just as for the standard phase contrast model above, the issue of uniqueness/stability reduces to inversion of the ray transform on $\mathcal{M}_{S_0}(\Omega_0)$.

General Inverse Scattering

Here we consider the case when the intensity in (37) is given by the general scattering operator in (6). Since deconvolution of the detector response does not impair upon uniqueness, we can without loss of generality assume that $\mathrm{PSF}^{\mathrm{det}} = \delta_{\boldsymbol{\omega}^\perp}$.

Uniqueness

As already argued for, we cannot expect uniqueness from local data, so we first assume U_{true} is supported in Ω_0. Next, non-uniqueness from phase retrieval can be resolved by making two measurements, or assuming the imaginary and real parts of the complex valued scatterer are related (section "Phase Retrieval"). Hence, a reasonable conjecture is that uniqueness holds for (50) under the same conditions as for the standard phase contrast model (section on p. 979). More precisely, U_{true} in (50) can be recovered uniquely whenever the real and imaginary parts of U_{true} are related as in (7) and micrographs are acquired at nonzero defocus (for a single thin lens section on p. 960) or at fixed distance form the specimen exit plane (for lens-less imaging section on p. 960). These assumptions are needed for phase retrieval (section "Phase Retrieval"), so another formulation is that U_{true} is uniquely determined from scattering data $\psi_{\mathrm{sc}}(\boldsymbol{\omega}, \cdot) = \mathcal{T}^{\mathrm{sc}}_{\boldsymbol{\omega}}(U_{\mathrm{true}})$ for $\boldsymbol{\omega} \in S_0$ whenever S_0 is infinite.

What mathematical results support the above conjecture? For full scattering data, there are uniqueness results for determining the far field pattern for (5) from a single measurement [35, theorem 10] (see chapter ▸ Inverse Scattering). One can also prove uniqueness for the inverse scattering problem with the magnetic Schrödinger operator in a slab geometry from a single scattered wave field, but this requires full knowledge of the Dirichlet-to-Neumann operator [104]. None of these results apply to our setting of tomographic data. The result closest to our setting is perhaps the one in [66] where one proves uniqueness for the inverse scattering problem when measurements are made on the reflected (backscattered) wave. For phase-less scattering data, [88] clearly states that uniqueness cannot hold for the problem of recovering U_{true} from the modulus of the far field pattern for (5) associated with a single direction $\boldsymbol{\omega}$. Here, Ω is assumed to be a bounded domain, and the scattered wave fulfills (4c) and an impedance boundary condition at $\partial\Omega$. A similar setting is studied in [95, 96], but this time data is the modulus of the scattered wave ψ_{sc} (that solves the direct scattering problem) acquired using wavelengths that vary continuously in an interval. A uniqueness result that is closest to our setting is the one in [90], which considers (5) with lens-less imaging. It turns out that for uniqueness one needs two micrographs for each direction, whereas stability requires four images. The results are however only proved when $S_0 \subset S^2$ is the equator and the scattering operator $\mathcal{T}^{\mathrm{sc}}_{\boldsymbol{\omega}}$ is approximated by the first order Born (paragraph on p. 956) or the paraboloid approximations. Nevertheless, it is highly probable that

the same result also holds when $\mathcal{T}_{\boldsymbol{\omega}}^{\mathrm{sc}}$ is not approximated and when S_0 an infinite set. Furthermore, if the imaginary part of the scatterer is related to the real part, one measurement should be enough for formal uniqueness.

Stability

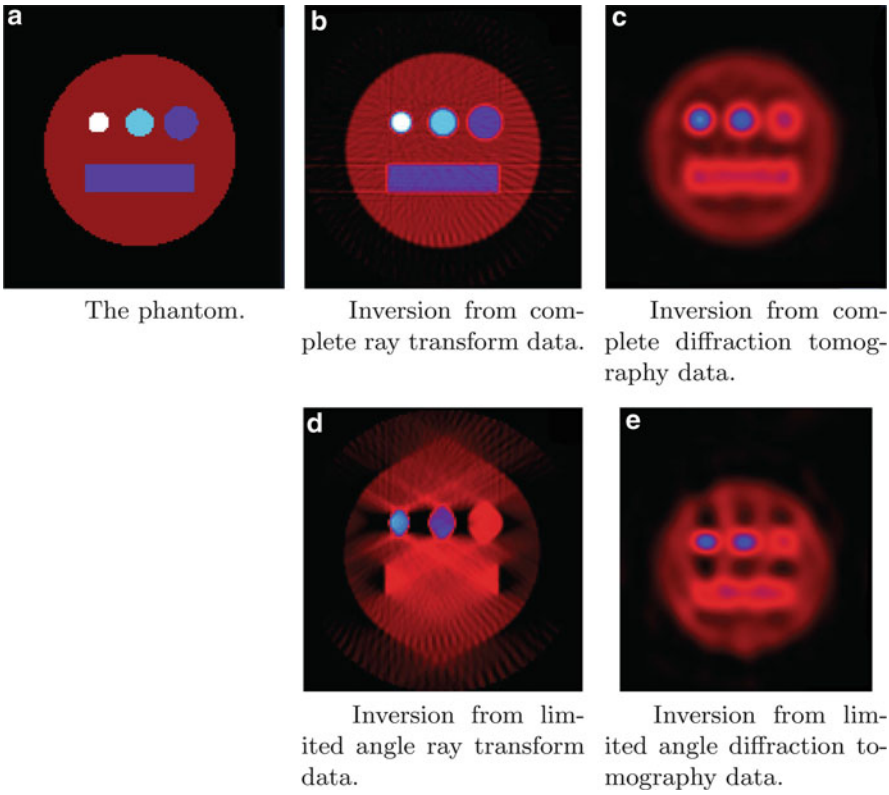
All results cited here deal with stability in recovering U_{true} from full scattering data $\psi_{\mathrm{sc}}(\boldsymbol{\omega}, \cdot) = \mathcal{T}_{\boldsymbol{\omega}}^{\mathrm{sc}}(U_{\mathrm{true}})$ for $\boldsymbol{\omega} \in S_0$.

There are various stability results for (4a) that assume full knowledge of the Dirichlet-to-Neumann map. Results closer to our setting consider the inverse scattering problem (5) with boundary conditions (4b)–(4c), $S_0 = S^2$, and Ω a bounded domain with smooth boundary. One can, e.g., show that stability increases as the wave number k increases [131, Theorem 1.1 and eq. (1.4)]. Similarly, [4] establishes Lipschitz stability for determining the low-frequency component of the potential. In [140] one considers the case when $U_{\mathrm{true}} = U_0 + \delta U$ where U_0 is known (typically representing some background) and δU is the (small) perturbation that is to be recovered. The result is that the perturbation can be stably reconstructed up to a certain frequency.

Stability in the context of phase contrast tomography is studied in [90], which considers (5) with lens-less imaging and $S_0 \subset S^2$ being the equator. They observe that numerical stability requires four measurements for each direction. A similar result is given in [130], which studies the same inverse scattering problem. If the real and imaginary parts of U_{true} are proportional (a special case of (7)) and two micrographs per direction are acquired, then the inversion is stable with respect to high-frequency noise. The arguments are however based on the projection assumption, but the result is interesting as it differs from situation in conventional tomography, which is known to be mildly ill-posed. This observation that phase contrast tomography is more stable than conventional tomography (assuming one can handle the phase problem) might also extend to the setting when $S_0 \subset S^2$ does not fulfill Orlov's criteria (section on p. 979). Hence, we expect wave imaging problems to be less influenced by limited angle data than ray imaging problems. The intuitive reason is that waves can "probe objects around corners" whereas rays cannot, and Fig. 5 illustrates this. There are however no formal mathematical results supporting this observation.

Nuisance Parameters

The formal statement of the inverse problem in ET in section "Mathematical Formulation" leaves out an important difficulty, namely the presence of nuisance parameters. These are parameters whose values need to be recovered alongside the potential U. We next list some of these parameters and indicate how they can be determined.



The phantom. Inversion from complete ray transform data. Inversion from complete diffraction tomography data.

Inversion from limited angle ray transform data. Inversion from limited angle diffraction tomography data.

Fig. 5 Sensitivity of wave and ray tomography in 2D against limited angle data (angular range is $\pm 60°$). Wave tomography is when data is restrictions of $\psi(\boldsymbol{\omega}, \cdot)$ for $\boldsymbol{\omega} \in S_0$ to a line orthogonal to $\boldsymbol{\omega}$ and where $\psi(\boldsymbol{\omega}, \cdot)$ solves (5) with (4b)–(4c). The wavelength of the incident wave equals the diameter of small white disc in the phantom in (**a**). It is clear that the limited angle artifacts (*bottom row*) are much more severe for ray tomography than for wave tomography. (**a**) The phantom. (**b**) Inversion from complete ray transform data. (**c**) Inversion from complete diffraction tomography data. (**d**) Inversion from limited angle ray transform data. (**e**) Inversion from limited angle diffraction tomography data (images courtesy of Frank Natterer)

Detector Parameters

The detector model outlined in (29) involves a detector response $\mathrm{PSF}^{\mathrm{det}}$, the overall gain C_{gain}, and the distribution of the background signal. None of these are nuisance parameters in the true sense since they can be determined by separate calibration experiments in a way independent of the tilt-series.

The overall gain C_{gain} can be calculated from a detector specification and/or from off-line data that has undergone basic pre-processing (section "Basic Pre-processing"), see, e.g., [146, pp. 380–381]. The detector response $\mathrm{PSF}^{\mathrm{det}}$ is rotationally symmetric and positive since the spreading of the signal due to the scintillator and fiber-optic/lens coupling is rotationally symmetric. Thus, $\mathrm{PSF}^{\mathrm{det}}$ is fully characterized by the Modulation Transfer Function (MTF), which is the modulus of its

Fourier transform: $\mathcal{F}_{\boldsymbol{\omega}^\perp}\big[\mathrm{PSF}^{\mathrm{det}}\big](\boldsymbol{\xi}) = \mathrm{MTF}\big(|\boldsymbol{\xi}|\big)$, Here, MTF: $\mathbb{R} \to \mathbb{R}_+$ measures how the signal amplitude is transferred for different spatial frequencies, and for slow-scan CCD sensors it is commonly parametrized as [176, p. 279], [146, p. 394]:

$$\mathrm{MTF}(q) := \frac{a}{1+\alpha q^2} + \frac{b}{1+\beta q^2} + c. \tag{56}$$

The above parameters can be estimated from off-line calibration measurements, see, e.g., [188], which in turn yields the detector response $\mathrm{PSF}^{\mathrm{det}}$. Finally, the background signal can also be estimated from similar calibration measurements.

Illumination and Optics Parameters

These are parameters associated with the illumination and optics. They are independent of the specimen but they can be unique for a tilt-series. Some, like the wavenumber k and magnification M, are considered constant and known to sufficient degree of accuracy. On the other hand, the image dose and a number of electron–optical parameters have nominal values that must be adjusted, either by an analysis of the recorded micrographs and/or by performing additional measurements after the ET data collection.

Image Dose

The image dose $N_0(\boldsymbol{\omega})$ in (37) plays an important role in the interpretation of data since it enters into the forward models in (41) and (49). It gives the number of electrons per pixel in the absence of a specimen.

One way to estimate the image dose is based on assuming a constant electron–current density J_0 (number of electrons per unit area and unit time). This quantity can be calculated by a Faraday cage and converted into the number of electrons per second. From the exposure time $t(\boldsymbol{\omega})$ we then get the actual dose used for the imaging by the relation $N_0(\boldsymbol{\omega}) = J_0\, t(\boldsymbol{\omega})$. Another option is to set $N_0(\boldsymbol{\omega}) = C_{\max}(\boldsymbol{\omega})/C_{\mathrm{gain}}$ where C_{gain} is the overall gain (see section on p. 982 for how it is estimated) and $C_{\max}(\boldsymbol{\omega})$ is the maximum pixel value in the micrograph. This will always underestimate the image dose. To account for the specimen, consider a linear absorption model where the specimen is a continuum medium in the slab Ω with thickness h. Hence, if $C(\boldsymbol{\omega})$ is the average detector response, then $C(\boldsymbol{\omega}) = C_0(\boldsymbol{\omega})\big(1-\mu h(\boldsymbol{\omega})\big)$ in which $C_0(\boldsymbol{\omega})$ is the image dose, μ is characterized by the specimen, and $h(\boldsymbol{\omega}) > 0$ is the thickness along $\boldsymbol{\omega}$ (section "Forward Operator for Amplitude Contrast Only"). If the image dose is the same, say C_0, for two images, then it is easy to see that it is given by $C_0 = \big(N_0(\boldsymbol{\omega}) - N_0(\boldsymbol{\omega}')\big)/\big(2-2h(\boldsymbol{\omega})/h(\boldsymbol{\omega}')\big)$. A more general approach is to use least-squares to fit a quadratic function to points on $\boldsymbol{\omega} \mapsto C_{\max}(\boldsymbol{\omega})$. This uses all data in the tilt-series and also allows one to estimate the shape of the specimen boundaries (a slab is the simplest shape but a "meniscus" type of membrane is more natural).

Remark 15. In experimental protocols, one typically reports the total dose at the object plane Γ_{obj} (Fig. 2). It roughly corresponds to the number of electrons per unit area at the specimen input plane Γ_{in}. This is an important number as it directly relates to the dose sensitivity of the specimen. The corresponding image dose is obtained by first dividing with the number of micrographs in the tilt-series, and then dividing by the magnification.

Optics Parameters

These parameters are only relevant for reconstruction methods based on forward models that account for the TEM optics, like the standard phase contrast model (section on p. 967).

First, we have parameters that describe the set-up for the optics (section "The Optical Set-Up") illustrated in Fig. 2. This set-up does not correspond to a physical optical system. The parameters f, p, and q can be selected freely as long as (20) holds. It is however wise to use the convention that f in (20) equals the focal length of the actual objective lens. The two relations will then give the values for p and q. An advantage of this convention is that the size of the aperture in the focal plane Γ_{focal} matches the size of the physical objective aperture. This convention provides the parameter $d := p - q$ associated with lens-free imaging, the simplest model for the optics (section on p. 960).

Next, a more advanced TEM optics model, like the one in section on p. 960, requires values for electron–optical parameters (CTF estimation). These enter into the definition of PSF^{op} in (24) with defocus being the most important one. In ET applications that do not require interpretation of 3D features smaller than 4 nm, there is no need for CTF estimation since nominal values of the electron–optical parameters are good enough. This is however not the case for high-magnification/resolution applications [57]. In general, CTF estimation is based on identifying shape and position of the so-called Thon rings in the power spectrum density. These are resolution-dependent amplitude modulations caused by the optics PSF [79, section 77.2]. This identification is an ill-posed problem [175] and a variety of estimation techniques have been developed, mainly for single particle applications, see [60, sections 2.37–2.38] and [51, section 8.5] for a more detailed overview. Unfortunately, most of these approaches are not applicable to ET since micrographs have too low contrast and/or are too noisy. One approach is based on strip-based periodogram averaging, extended throughout the tilt-series to overcome the low contrast conditions found in ET [57, 191]. Another alternative, that better respects the true Poisson characteristics of the noise, is outlined in [186].

Remark 16. For material sciences applications, CTF estimation is important due to the high resolution/magnification. The estimation problem is in part simpler since micrographs have better contrast and signal-to-noise ratio, but more CTF parameters

have to be estimated as well. The reader may consult [176, chapter 10] for a survey of approaches.

Specimen-Dependent Parameters

In ET there are also parameters that depend on the specimen. One such parameter relates to local data (p. 974). Since the support of U is not contained in $\Omega_0 \subset \Omega \subset \mathbb{R}^3$, pixels in all micrographs (besides the zero-tilt image) will have contribution from outside Ω_0 (local tomography artifacts), see, e.g., [123, section 4]. These artifacts become more pronounced when there is strongly scattering material outside Ω_0, as for biological in situ specimens and material sciences. A typical approach is to "pad" the data, which here means removing contributions from Ω_0 (long-object compensation). The simplest padding is based on assuming U equals some constant average value outside Ω_0 (constant continuation), which then becomes a specimen-dependent parameter. One can also consider smooth continuation in order to avoid introducing edge effects, but that results in introducing further specimen-dependent parameters. Yet another approach is to consider a low-resolution reconstruction from a larger region of interest, and use that to pad data.

Another parameter is the scaling factor Q associated with the assumption in (7) that relates the real and imaginary part of the potential. Current methods for estimating the amplitude contrast ratio are based on performing additional measurements, such as taking a defocus series. This is unfortunately not possible for most biological specimens.

8 Data Pre-processing

This section describes data processing steps that are commonly applied to micrographs in a tilt-series before they are used within a reconstruction method.

Basic Pre-processing

The degree of pre-processing of micrographs depends on how much one seeks to include in the forward operator. For all ET applications it is highly recommended to work with data that has been flat-fielded. Flat-fielding refers to the process of compensating for different gains and dark currents in a detector, so a uniform signal corresponds to a uniform output. This process also includes gain normalization that removes image features associated with the structure of the scintillator crystal and fiber-couplings (section on p. 963), see an example of this effect in [146, p. 393]. In addition, it is customary to remove high frequency peaks (X-ray peaks) based on a local mean value calculations.

Alignment

Acquiring a tilt-series corresponds to sampling a real valued function defined on $\mathcal{M}_{S_0}(\Omega_0)$ (Definition 1). The nominal specification for this sampling is unfortunately not accurate enough since the specimen undergoes unintentional movements during in between the acquisition of the micrographs. These movements result in shifts, rotations, and other types of image distortions.

Alignment is the procedure that corrects for these distortions and determines how $\mathcal{M}_{S_0}(\Omega_0)$ is sampled. Most alignment procedures follow a common four-step scheme:

Initial processing: In addition to basic pre-processing (section "Basic Pre-processing"), micrographs in tilt-series often also undergo further processing that involves denoinsing and feature extraction specifically tailored for alignment. These micrographs are only used during the alignment; to avoid losing information, they are not used for reconstruction.

Coarse alignment: Despite high-precision mechanics, raw tilt series might contain large shifts and rotations. Cross-correlation is used to coarsely align adjacent images. After coarse alignment, we can assume a smooth trajectory of features across images, facilitating subsequent alignment steps.

Feature tracking: The most common technique for feature detection is to define a patch around the feature as a template, and cross-correlate this template with other micrographs to search for the feature of interest. Due to high noise in images, one often have many false positive detections.

Determine the geometric relation: Once the 2D location of each feature is established and tracked across multiple images, epipolar geometry can be used to establish the geometric relation between the various images.

The choice of alignment method is based on the type of feature one seeks to track. Cryo-fixated specimens, especially in vitro specimens, lack visible features. Hence, one has to add fiducial markers to align tilt-series of such specimens. These are spherical gold beads that appear as high contrast point-like features in the micrographs. Techniques from epipolar geometry can then be used to elucidate the geometric relation between the micrographs, see [3] and the references therein for more details.

Deconvolving Detector Response

The Fourier transform of the detector response is strictly positive (section on p. 982), so it is fairly straightforward to deconvolve it, e.g., using a Wiener filter or the Richardson–Lucy algorithm [197]. A benefit of deconvolving the detector response, instead of including it into the forward model, is that the noise model for (deconvolved) data (section on p. 964) becomes much simpler.

Deconvolving Optics PSF

Deconvolving the optics response is only relevant for tilt-series from HRTEM imaging where phase contrast is important.

For lens-free imaging (section on p. 960), this corresponds to deconvolving the effect of free-space propagation. Several approaches exist for this within the X-ray phase contrast imaging community, see [22] for a survey.

Optics models that are based on the single thin lens with an aperture model (section on p. 960) are more involved to deconvolve since the CTF (the Fourier transform of the PSF) in (24) has multiple zeros. As an example, for the standard phase contrast model (section "Standard Phase Contrast Model"), one has to deconvolve $\mathrm{PSF}^{\mathrm{op}}_{\mathrm{tot}}$ given in (45). If $Q(\boldsymbol{\omega}) = 0$, then $\mathrm{PSF}^{\mathrm{op}}_{\mathrm{tot}}(\mathbf{0}) = \mathrm{PSF}^{\mathrm{op}}_{\mathrm{im}}(\mathbf{0}) = 0$, i.e., low-frequency image information is lost. Hence, the optics deconvolution is more challenging for HRTEM imaging specimens that act as "pure phase" objects. In general, the challenge in regularizing the optics deconvolution is the difficulty in specifying reasonable a priori information about the "true" signal, which here would be the micrograph one would obtain without perfect optics.

The optics is on the other hand needed to make phase contrast visible, so its deconvolution is only sensible if it is coupled with phase retrieval (section "Phase Retrieval"). In conclusion, it is highly dubious to first deconvolve the optics and then try 3D reconstruction if one does not couple that with some kind of phase retrieval. Thus, one either ignores the optics or include it into the forward model, so optics deconvolution is performed simultaneously with 3D reconstruction.

Phase Retrieval

For each direction $\boldsymbol{\omega} \in S_0$, we have a phase retrieval problem (also called exit-wave reconstruction problem) which amounts to inverting the intensity operator in (26). In general this problem does not have a unique solution.

A common approach is to record two, or more, micrographs of the specimen using different imaging conditions (like different defocus values). From such data one can then reconstruct the phase of the wave [58]. This principle is put into action in electron holography and approaches based on the transport-of-intensity equation [99]. None of these approaches are however applicable to ET in life sciences since they require multiple micrographs for each direction $\boldsymbol{\omega} \in S_0$, which is often not possible due to the specimen dose sensitivity (section "The Dose Problem"). Thus, we are left with considering "in-line" approaches that use only one micrograph at each direction.

Let us first consider the simplest model for the optics (lens-less imaging, section on p. 960), i.e., free-space propagation. In this setting, phase retrieval within ET is identical to the corresponding problem in in-line X-ray phase contrast tomography. A variety of methods have been developed for the former, all based on relating the real and imaginary parts of the complex refractive index as in (7). The idea is to make use of the optics operator $\mathcal{T}^{\mathrm{op}}$, which in this case is given by free-space

propagation (21), and perform the phase retrieval and optics deconvolution in one step by inverting $\mathcal{I} \circ \mathcal{T}^{\text{op}}$. The resulting phase retrieval operation is expressible as a convolution, see [22] for a nice survey. For ET, there is no obvious relation between the electrostatic and the absorption potentials. Nevertheless, it is common to assume that the latter is a multiple times the former, as discussed in section on p. 949. Thus, the aforementioned approaches from in-line X-ray phase contrast tomography should be applicable also on ET data as a pre-processing step for phase retrieval. A similar approach should also be applicable when the optics model is more complex, such as in (23).

Finally, one also has the option to include the intensity into the forward operator. In that context, phase retrieval is performed jointly as part of 3D reconstruction. If one seeks an affine forward operator, then one must linearize the intensity as in (26).

9 Reconstruction Methods

Before we describe the various reconstruction methods for ET, we state assumptions and procedures that are common for all current methods in use.

Type of Forward Operator

Unless otherwise stated, all reconstruction methods assume the forward operator is given by either the amplitude contrast (section "Forward Operator for Amplitude Contrast Only") or the standard phase contrast (section on p. 967) models. For data that has undergone basic pre-processing (section "Basic Pre-processing"), both these lead to affine forward operators expressible as

$$\mathcal{T}(U)(\boldsymbol{\omega}, \boldsymbol{x}) = C(\boldsymbol{\omega}) - C_0 \Big\{ \mathrm{PSF}(\boldsymbol{\omega}, \cdot) \underset{\boldsymbol{\omega}^\perp}{\circledast} \mathcal{P}(U_{\text{re}})(\boldsymbol{\omega}, \cdot) \Big\}(\boldsymbol{x}) \quad \text{for } \boldsymbol{x} \in \boldsymbol{\omega}^\perp. \tag{57}$$

The specific expression for the PSF depends on the model but the overall structure remains the same for the amplitude and standard phase contrast models. In this context, the term $C(\boldsymbol{\omega})$ represents data one would record if there would be no specimen. It can be, together with C_0, estimated from data (section "Nuisance Parameters"), in which case the inverse problem in ET can be recast to a setting with a linear forward operator.

Local Data

In ET we have local data (see p. 974), i.e., the support of U_{true} is not contained in Ω_0. Hence, each micrograph contains the contribution of a larger or smaller extent of the specimen, so (53) becomes inconsistent. This is not an issue for reconstruction methods that are not sensitive to such issues, like ELT (section on p. 994). Other reconstruction methods should either pre-process data as to remove contributions from outside Ω_0, or adapt the evaluation of the forward operator and its adjoint to account for this contribution (section on p. 985). This is referred to as long-object compensation and is based on assumptions about U_{true} outside Ω_0, see [192] for

artifacts that arise in ET when such compensation is not performed in iterative methods (section "Iterative Methods with Early Stopping").

Handling Nuisance Parameters

A reconstruction method for ET must handle how to assign values to nuisance parameters (section "Nuisance Parameters"), like $\boldsymbol{\omega} \mapsto C(\boldsymbol{\omega})$ and C_0 in (57). Unless otherwise stated, all reconstruction methods in ET assume these are determined before reconstruction.

One can in principle consider the nuisance parameters as part of the signal that is to be recovered and attempt at reconstructing these alongside the initial signal. This is theoretically possible when using iterative and variational approaches. On the other hand, the forward operator might be nonlinear w.r.t. the nuisance parameters, the stability of the inversion w.r.t. the nuisance parameters could vary considerably when compared to inversion of the original signal, it is unclear which data discrepancy to use, etc. Hence, in practice one adopts an intertwined approach in which the signal is first reconstructed keeping the nuisance fixed, then the nuisance parameters are recovered keeping the signal fixed [165].

Type of Mathematical Results

From a strict mathematical point of view, a regularization involves a reconstruction operator that is well-posed (unique and stable solution) when reconstruction parameter(s) are chosen accordingly. Next, regularized solutions must converge to a ML/least-squares solution as the data error $\|\boldsymbol{g}^{\text{noise}}\| \to 0$. Besides these two requirements, one often also studies convergence rates (estimate of the difference between the regularized solution and a ML/least-squares solution) and stability estimates (bounds to the difference between the regularized solution with noise-free data and noisy data).

Analytic Methods

Overview

Analytic methods are applicable to inverse problems where the forward operator has an inverse that has an analytic formula, which is a concatenation of "simple" mathematical operators (such as finite number of differentiations to any order and integrations, whereas analytic continuation or summation of infinite series is not simple). The starting point is the reconstruction problem with continuum data given in (50). Assuming this problem has a unique solution, one can define the *reconstruction operator* as the inverse of the forward operator (*exact reconstruction*). On the other hand, if there is instability and/or non-uniqueness, like in ET (section "Incomplete Data, Uniqueness and Stability"), then an appropriate reconstruction operator should recover a feature, such as an approximate identity, of the signal. The actual implementation is given by an appropriate discretization of the reconstruction operator.

In summary, analytical methods lead to algorithms in which the signal is directly calculated from the measurements in a single step, without resorting to more time-

consuming, iterative methods. In tomographic applications, the numerical analysis basically employs filtering, backprojection, and summation operations, as well as discrete Fourier transforms.

Backprojection-Based Methods

The Filtered Back-Projection (FBP) and Weighted Back-Projection (WBP) methods are examples of backprojection-based methods that constitute the standard approach within the ET community for solving the reconstruction problem, see [143] for a recent survey.

Data Pre-processing

All backprojection-based methods assume data is (noisy) samples of the ray transform, i.e., the forward operator is (57) with $C(\boldsymbol{\omega}) = 0$, $C_0 = 1$, and $\mathrm{PSF}(\boldsymbol{\omega}, \cdot) = \delta_{\boldsymbol{\omega}^\perp}$. Hence, data should be pre-processed so that it fits this assumption, i.e., nuisance parameters $C(\boldsymbol{\omega})$ and C_0 need to be estimated and the PSF needs to be deconvolved.

A Priori Information

The assumption is that noise is high frequency whereas relevant features of U_{true} are low frequency. Thus, by providing an approximate inverse, these relevant features of U_{true} are recovered while noise is suppressed.

Reconstruction Operator

The starting point for both FBP and WBP is an equality that relates the ray transform to its back-projection. Let $S_0 \subset S^2$ be an infinite measurable set and $U \in \mathscr{S}(\mathbb{R}^3, \mathbb{R})$. A simple modification to the arguments used in proving [136, eq. (2.34)] allows us to prove the following identity on $\mathcal{M}_{S_0}(\Omega_0)$:

$$U * H = \mathcal{P}^*_{S_0}\big(h \underset{\boldsymbol{\omega}^\perp}{\circledast} \mathcal{P}(U)\big) \quad \text{for } h \in \mathscr{S}\big(\mathcal{M}_{S_0}(\Omega_0), \mathbb{R}\big) \text{ and } H := \mathcal{P}^*_{S_0}(h). \tag{58}$$

Here, $\mathcal{P}$ is the ray transform in (1) and $\mathcal{P}^*_{S_0}$ is the *(parallel beam) back-projection* defined as follows:

Definition 2 (Back-Projection). Let $g \in \mathscr{S}\big(T(S^2), \mathbb{R}\big)$ and $S_0 \subset S^2$ is a measurable set. The *(parallel beam) back-projection* of g is then defined as

$$\mathcal{P}^*_{S_0}(g)(\boldsymbol{x}) := \int_{S_0} g\big(\boldsymbol{\omega}, \boldsymbol{x} - (\boldsymbol{x} \cdot \boldsymbol{\omega})\boldsymbol{\omega}\big)\, \mathrm{d}\boldsymbol{\omega} \quad \text{for } \boldsymbol{x} \in \mathbb{R}^3, \tag{59}$$

with $\mathrm{d}\boldsymbol{\omega}$ denoting the surface measure on S_0.

Remark 17. $\mathscr{S}\big(T(S^2), \mathbb{R}\big)$ is the Schwarz space for real-valued functions defined on $T(S^2)$, see [136, section 2.2] for the formal definition. $\mathcal{P}^*_{S_0}$ maps such functions to functions on $\mathbb{R}^3$. Furthermore, if the $\mathscr{L}^2$-inner product is considered for both the domain and range of $\mathcal{P}^*_{S_0}$, then $\mathcal{P}^*_{S_0}$ is the adjoint of $\mathcal{P}$.

The reconstruction operator for the FBP method is now defined as the right-hand side of the equality in (58):

$$\mathcal{R}_{\mathrm{FBP}}(g) := \mathcal{P}_{S_0}^*\big(h \underset{\omega^\perp}{\circledast} g\big) \quad \text{for } g \in \mathscr{S}\big(\mathcal{M}_{S_0}, \mathbb{R}\big). \tag{60}$$

Hence, by (58) we get that $\mathcal{R}_{\mathrm{FBP}}(g) = U_{\mathrm{true}} * H$ whenever $g = \mathcal{P}(U_{\mathrm{true}})$ on $\mathcal{M}_{S_0}(\Omega_0)$. In this context, $h\colon \mathcal{M}_{S_0} \to \mathbb{R}$ is called the *reconstruction kernel* and H is the corresponding *filter*. The idea in FBP is now to choose h such that $H \approx \delta$, since this implies that $\mathcal{R}_{\mathrm{FBP}}(g) \approx U_{\mathrm{true}}$.

The *WBP* method is mathematically equivalent to FBP. It is based on taking $h = \delta_{\omega^\perp}$ in (58). This means, of course, that the corresponding filter H will not be an approximation of δ, so (60) does not directly yield a reconstruction of U_{true}. However, if one can derive an expression for the Fourier transform of $H \approx \mathcal{P}_{S_0}^*(\delta_{\omega^\perp})$, then one can use this to recover U_{true} by dividing the Fourier transform of $\mathcal{P}_{S_0}^*(g)$ by the Fourier transform of H, and then apply the inverse Fourier transform. This leads to the following reconstruction operator:

$$\mathcal{R}_{\mathrm{WBP}}(g) := \mathcal{F}^{-1}\left[\frac{\mathcal{F}\left[\mathcal{P}_{S_0}^*(g)\right]}{\mathcal{F}\left[H\right]}\right].$$

For FBP, in order to get a useful reconstruction operator one must choose the reconstruction kernel h such that convolution with $\mathcal{P}_{S_0}^*(h)$ represents extracting some feature of U. The idea is to first derive the reconstruction kernel h for exact reconstruction from ideal data, i.e., $\mathcal{P}_{S_0}^*(h) = \delta$. The regularized variant that provides an approximative inverse for noisy data is then given as a band-limited variant of the aforementioned reconstruction kernel. Similarly, in WBP we seek H such that it approximately equals the backprojection of a Dirac delta.

Adaptation to ET

The adaptation to ET lies in finding an appropriate reconstruction kernel h for the FBP method. The corresponding problem of finding H for WBP is discussed in [143, 156].

For single-axis tilting (section "Examples Relevant for ET"), the entire 3D ray transform inversion problem reduces to a sequence of 2D inversion problems that can be handled separately. Hence, in that sense, single-axis tilting does not lead to a truly 3D inversion problem. Likewise, the current approach for FBP/WBP reconstruction from double-axis tilting data is to split such a data set into two single-axis data sets, perform 3D reconstruction for each of these, and then combine the two 3D reconstructions by some averaging. This might give the impression that reconstruction from double-axis tilting is also not a "true 3D problem". This is however not the case.The two 3D reconstructions will have different limited angle artifacts, so it is clear that a spatially dependent averaging will perform better than just a regular average. One can derive a partition of unity to use for such a spatially

dependent averaging by considering the reconstruction kernel for the 3D problem in double axis tilting.

Let us now consider the reconstruction kernel for general parallel beam line complexes in which S_0 is a curve. To derive an equation for the reconstruction kernel (filter equation), we first consider the following relation [136, theorem 2.17]:

$$\mathcal{F}[H](\boldsymbol{\xi}) = \frac{1}{|\boldsymbol{\xi}|}\int_{S_0\cap\boldsymbol{\xi}^\perp} \mathcal{F}_{\boldsymbol{\omega}^\perp}[h(\boldsymbol{\omega},\cdot)](\boldsymbol{\xi})\,\mathrm{d}\boldsymbol{\omega} \quad \text{for } h \in \mathscr{S}\big(\mathcal{M}_{S_0},\mathbb{R}\big),$$

where $H := \mathcal{P}^*_{S_0}(h)$. Exact reconstruction requires $H = \delta$, i.e., $\mathcal{P}^*_{S_0}(h) \equiv 1$, so

$$\int_{S_0\cap\boldsymbol{\xi}^\perp} \mathcal{F}_{\boldsymbol{\omega}^\perp}[h(\boldsymbol{\omega},\cdot)](\boldsymbol{\xi})\,\mathrm{d}\boldsymbol{\omega} = |\boldsymbol{\xi}|.$$

Hence, an obvious choice of reconstruction filter h, independent of $\boldsymbol{\omega}$, is

$$\mathcal{F}_{\boldsymbol{\omega}^\perp}[h(\boldsymbol{\omega},\cdot)](\boldsymbol{\xi}) := \frac{|\boldsymbol{\xi}|}{\big|S_0\cap\boldsymbol{\xi}^\perp\big|} \quad \text{where} \quad \big|S_0\cap\boldsymbol{\xi}^\perp\big| := \int_{S_0}\delta(\boldsymbol{\omega}\cdot\boldsymbol{\xi})\,\mathrm{d}\boldsymbol{\omega}. \tag{61}$$

This is fine as long as S_0 satisfies Orlov's criteria (section "Incomplete Data, Uniqueness, and Stability"), which ensures that $\big|S_0\cap\boldsymbol{\xi}^\perp\big| \neq 0$. Unfortunately, in ET this is not the case. A natural modification to (61) is to set the Fourier transform of $h(\boldsymbol{\omega},\cdot)$ to zero at points $\boldsymbol{\xi}$ where $\big|S_0\cap\boldsymbol{\xi}^\perp\big| = 0$. With this choice of reconstruction filter, the FBP reconstruction operator in (60) provides a least-squares solution with minimal 2-norm instead of an exact solution.

To get a more explicit expression for h, let A_{S_0} be the set of all $\boldsymbol{\xi} \neq 0$ in $\mathbb{R}^n$ such that $\boldsymbol{\xi}^\perp$ is not tangent to S_0 and intersects it at a finite number of points, with $N_{\boldsymbol{\xi}}$ denoting the number of such intersection points. Also, parametrize the curve S_0 by $s \mapsto \boldsymbol{\gamma}(s)$ where $s \in I$ is an interval and $\mathrm{d}\boldsymbol{\omega} = \big|\boldsymbol{\gamma}(s)\big|\,\mathrm{d}s$. From a direct modification of the proof of [136, theorem 2.17], we get

$$\mathcal{F}[H](\boldsymbol{\xi}) = 2\pi\sum_{i=0}^{N_{\boldsymbol{\xi}}}\frac{\big|\boldsymbol{\gamma}(s_i)\big|}{\big|\boldsymbol{\xi}\cdot\dot{\boldsymbol{\gamma}}(s_i)\big|}\mathcal{F}_{\boldsymbol{\omega}^\perp}\big[h\big(\boldsymbol{\gamma}(s_i),\cdot\big)\big](\boldsymbol{\xi}) \quad \text{for } \boldsymbol{\xi} \in A_{S_0},$$

where $s_1,\ldots,s_{N_{\boldsymbol{\xi}}} \in I$ are such that $\boldsymbol{\xi}\cdot\boldsymbol{\gamma}(s) = 0$. The above immediately gives us a necessary condition for h (filter equation) if $H = \delta$ is to hold:

$$2\pi\sum_{i=0}^{N_{\boldsymbol{\xi}}}\frac{\big|\boldsymbol{\gamma}(s_i)\big|}{\big|\boldsymbol{\xi}\cdot\dot{\boldsymbol{\gamma}}(s_i)\big|}\mathcal{F}_{\boldsymbol{\omega}^\perp}\big[h\big(\boldsymbol{\gamma}(s_i),\cdot\big)\big](\boldsymbol{\xi}) = 1 \quad \text{for all } \boldsymbol{\xi} \in A_{S_0}. \tag{62}$$

Thus, a natural choice of h is to define it as the inverse Fourier transform of

$$\boldsymbol{\xi} \mapsto \begin{cases} \dfrac{\left|\boldsymbol{\xi} \cdot \dot{\boldsymbol{\gamma}}(s)\right|}{2\pi N_{\boldsymbol{\xi}} \left|\boldsymbol{\gamma}(s)\right|} & \text{whenever } N_{\boldsymbol{\xi}} \neq 0, \\ 0 & \text{otherwise.} \end{cases} \tag{63}$$

Note that there are multiple choices for h whenever $N_{\boldsymbol{\xi}} > 1$ for some $\boldsymbol{\xi}$. Furthermore, the filter equation (62) is a necessary condition for $H = \delta$, so an interesting question is to find sufficient conditions.

Remark 18. There is a great deal of literature in the tomography community for constructing filter equations, i.e., designing reconstruction kernels h for 3D FBP methods such that $H = \delta_0$. Most of these results are not applicable to ET since they deal with data collection geometries (cone-beam or helical cone-beam) that arise in medical X-ray tomography and therefore are not parallel beam. Some work on filter equations in the parallel beam context can be found in early work on positron emission tomography [42]. These results are not applicable to ET since they assume $S_0 \subset S^2$ is an open set.

Remark 19. In the 3D electron microscopy literature there are several publications devoted to deriving reconstruction kernels when S_0 is not a curve, but rather some irregular subset of S^2. This is motivated from the needs in single particle analysis when one seeks to invert the ray transform on a parallel beam line complex in which S_0 is an open set and the sampling is irregular/random within S_0. These results are nicely surveyed in [143].

Comments and Discussion

The main advantage of FBP/WBP is that it is fairly easy to implement, so most software packages for ET have at least one of these methods implemented as listed in [143]. The methods are furthermore efficient enough to allow for large scale 3D reconstructions in short time. Next, the reconstruction quality is sufficient for answering biological questions that only require interpretation of 3D structural details larger than 60–80 nm.

There are several disadvantages that comes with using FBP/WBP. First, one assumes a forward model given by the ray transform, so data should be pre-processed accordingly to reflect that assumption. Next is to choose a reconstruction kernel to yield an approximative inverse, which in FBP is governed by the band-limit that becomes the regularization parameter. Techniques from sampling theory can now be used to identify efficient sampling schemes, provide qualitative understanding of certain artifacts, and provide guidelines for how to band-limit given a sampling scheme. This is well understood in the planar setting (i.e., in $\mathbb{R}^2$) for complete and moderately noisy fan or parallel beam data [52, 53, 134], but the situation is less clear in 3D and/or when data is incomplete. Guidelines, like Crowther's criterion [145, p. 316], that relate the "resolution" to the sampling scheme are therefore not applicable to ET even when data is collected by single-axis tilting. Finally, FBP/WBP are not really based on a model of the stochasticity

of the data. Both methods are essentially a discretization of an inversion formula. Therefore, it is difficult (if not impossible) to devise schemes for selecting the filter h that takes into account the specific stochasticity of the data, which in the ET setting is essentially Poisson distributed.

A common approach when using FBP/WBP in ET is to post-process reconstructions to remove speckle, enhance features, and reduce limited angle streak artifacts. The simplest approach for single-axis data is to average over slices. Most software packages offer this possibility to reduce the influence of noise. Other post-process approaches seek to reduce the well-known streak artifacts, see, e.g., [62, 133] in which the latter uses microlocal analysis to explain the appearance of the aforementioned artifacts.

Electron Λ-Tomography (ELT)

ET data are local so the corresponding inverse problem does not have a unique solution even for continuum noise-free data (section "Incomplete Data, Uniqueness, and Stability"). A natural question is to determine what features that are uniquely recoverable from such local data and to provide stable means for their recovery.

As we shall see, one such feature is the "visible singularities" that provide location of certain edges of objects in the specimen. The visibility depends on the data collection geometry. Furthermore, it turns out that the recovery of the visible singularities is only mildly ill-posed (comparable to inversion of the ray transform on complete data). ELT is a local reconstruction method that recovers these visible singularities that is based on Λ-tomography. The description given here is based on [151, 155], see also chapter ▸ Microlocal Analysis in Tomography.

Data Pre-processing

Same as for backprojection-based methods (section "Backprojection-Based Methods").

A Priori Information

Relevant edge information can be recovered by suppressing high-frequency phenomena.

Uniquely Recoverable Signal Features

The analysis of signal features that are uniquely recoverable from local tomographic data relies heavily on microlocal analysis. Microlocal analysis was initially introduced to study how singularities propagated in solutions of partial differential equations [86, 168], and its application to integral geometry was first given in [76]. Somewhat later, microlocal analysis was used in a series of papers for studying the admissibility problem for various restricted generalized ray transforms (integrating over families of geodesics) [73–75]. Further applications in integral geometry came with [20] (support theorems) and [150] (limited data problems for the ray transform in $\mathbb{R}^2$ and $\mathbb{R}^3$), see chapter ▸ Microlocal Analysis in Tomography for a more detailed description.

The central idea in microlocal analysis is that singularities are characterized not only by their location, but also by the high frequencies that cause them. To make this more precise, one makes use of the well-known Fourier characterization of smoothness, i.e., U is smooth if and only if its Fourier transform decays faster than any power of $1/|\boldsymbol{\xi}|$ as $|\boldsymbol{\xi}| \to \infty$. Localizing this characterization yields a description of the singular support (location of the singularity). The formal definition of a wavefront set is given by a further localization, microlocalization, that characterizes those directions in Fourier space where we lack smoothness (i.e., direction of those high frequencies causing the singularities):

Definition 3 (The Wavefront Set). Let $\boldsymbol{x}_0 \in \mathbb{R}^n$ and $\boldsymbol{\xi}_0 \in \mathbb{R}^n \setminus \{\mathbf{0}\}$. We say that U *is smooth near* $\boldsymbol{x}_0$ *in the direction* $\boldsymbol{\xi}_0$ if there is a smooth cut-off function ϕ near $\boldsymbol{x}_0$ (i.e. ϕ has compact support and $\phi(\boldsymbol{x}_0) \neq 0$) such that $\mathcal{F}[U\phi]$is rapidly decaying in some open conical neighborhood of $\boldsymbol{\xi}_0$. The *wave front set*, $\mathrm{WF}(U)$, of U are those points $(\boldsymbol{x}_0, \boldsymbol{\xi}_0)$ where U is not smooth near $\boldsymbol{x}_0$ in the direction $\boldsymbol{\xi}_0$.

A simple example might aid the intuitive understanding. If U has a jump discontinuity along a smooth hyper-surface $\Gamma \subset \mathbb{R}^n$, then $\mathrm{WF}(U)$ consists of $(\boldsymbol{x}_0, \boldsymbol{\xi}_0)$ where $\boldsymbol{x}_0 \in \Gamma$ and $\boldsymbol{\xi}_0$ is normal to Γ.

Remark 20. The above definition skips some mathematical technicalities, e.g., $\boldsymbol{\xi}_0$ is actually a cotangent vector and not an element in the same space as $\boldsymbol{x}_0$. This distinction becomes important when one defines the wavefront set for functions (or distributions) defined on smooth manifolds. Furthermore, one can also introduce Sobolev wavefront sets that encodes Sobolev regularity, see [151] and chapter ▸ Microlocal Analysis in Tomography for details.

Next, for inverse problems it is natural to seek to characterize singularities of the signal that are detectable in data. In the context of inverting the ray transform, a singularity $(\boldsymbol{x}, \boldsymbol{\xi}) \in \mathrm{WF}(U)$ is said to be *visible* from ray transform data $g = \mathcal{P}(U)$ if it corresponds to a singularity in $\mathrm{WF}(g)$. Thus, visible singularity are those that leave traces in data, and these can be characterized if data g is given on a parallel beam line complex:

Theorem 1 (Microlocal Regularity Principle). *A singularity* $(\boldsymbol{x}, \boldsymbol{\xi}) \in \mathrm{WF}(U)$ *is visible from ray transform data* $g = \mathcal{P}(U)$ *on* $\mathcal{M}_{S_0}(\Omega_0)$ *if and only if there is a line in* $\mathcal{M}_{S_0}(\Omega_0)$ *through* $\boldsymbol{x}$ *to which* $\boldsymbol{\xi}$ *is co-normal (with a few exceptions). Furthermore, the recovery of the visible singularities is mildly ill-posed in the sense that the singularities in the data g are weaker than those of U by 1/2 Sobolev order (good enough to allow stable detection in practice).*

An important special case is when U, as in the example following Definition 3, is smooth except for a jump discontinuity along a smooth hyper-surface Γ. Then, a singularity at $\boldsymbol{x} \in \Gamma$ is visible if and only if there is a line in $\mathcal{M}_{S_0}(\Omega_0)$ that goes through $\boldsymbol{x}$ and is tangent to $\mathcal{M}_{S_0}(\Omega_0)$ (with a few exceptions).

Remark 21. In order to simplify the presentation, we have left out an array of technicalities in the formulation of Theorem 1, see, e.g., [151, theorem 6.3] for a rigorous formulation in the single-axis tilting case. Note also that this theorem is not dealing with a specific reconstruction method. It is more analogous to a "uniqueness" result describing what is possible to recover (assuming one has no further a priori information) irrespective of the reconstruction method.

Reconstruction Operator

We have seen that visible singularities are signal features that are uniquely recoverable from continuum ray transform data on $\mathcal{M}_{S_0}(\Omega_0)$. ELT is a reconstruction method that recovers such singularities from ET data. It is an adaptation of Λ-tomography, which was developed independently of microlocal analysis and first introduced in [173, 178] as a local tomographic reconstruction method.

To describe ELT, we start off by considering $S_0 = S^2$. Then, [52, eqs. (4–6)] yields the following generalization of (58) for $m \geq -1$:

$$\Lambda^m(U) * H = \mathcal{P}^*_{S_0}\big(h \underset{\boldsymbol{\omega}^\perp}{\circledast} \mathcal{P}(U)\big) \quad \text{on } \mathcal{M}_{S_0}(\Omega_0) \text{ for } H := \mathcal{P}^*_{S_0}\Lambda^{-m}(h). \tag{64}$$

Here, $\mathcal{P}^*_{S_0}$ is the backprojection operator given in (59) and Λ is Calderón's operator that is defined in terms of the Fourier transforms as $\mathcal{F}[\Lambda(U)](\boldsymbol{\xi}) := |\boldsymbol{\xi}|\mathcal{F}[U](\boldsymbol{\xi})$ for $U \in \mathscr{S}(\mathbb{R}^3, \mathbb{C})$. Now, we proceed like in the FBP method, i.e., we first derive a reconstruction kernel h for exact recovery of $\Lambda^m(U)$ by selecting h so that $H = \delta$ in (64). The regularized variant is obtained by band-limiting the aforementioned reconstruction kernel, which in turn provides an approximative inverse of $\Lambda^m(U)$. The case $m = 1$ is of specific interest since $\Lambda(U)(\boldsymbol{x})$ is unaffected by local tomography artifacts (section on p. 985) as its calculation by (64) only involves lines passing through a small neighborhood of $\boldsymbol{x}$.

In ET, $S_0 \subset S^2$ is a curve. For simplicity, consider the case when S_0 is the equator in the (x, y)-plane. Then $\mathcal{P}^*_{S_0}\mathcal{P}U = k * U$ where

$$k(\boldsymbol{x}) := c\delta(z)\frac{1}{|\boldsymbol{x}'|} \quad \text{with } \boldsymbol{x}' = (x, y).$$

Let Λ' and $\triangle'$ denote the 2D variants of Λ and the Laplacian restricted to the (x, y)-plane, i.e., $\Lambda' := \sqrt{-\partial_1^2 - \partial_2^2}$ and $\triangle' := \partial_1^2 + \partial_2^2$. Then,

$$\Lambda'\mathcal{P}^*_{S_0}\mathcal{P}U = cU \quad \text{so} \quad \triangle'\,\mathcal{P}^*_{S_0}\mathcal{P}U = c\Lambda' U. \tag{65}$$

Since the tangent to the great circle $\omega_3 = 0$ is contained in ω_3-plane and is orthogonal to $\boldsymbol{\omega}$, we can write (65) as

$$\Lambda'\mathcal{P}^*_{S_0}\mathcal{P}U = \mathcal{P}^*_{S_0}\sqrt{-\mathcal{D}^2_{S_0}}\mathcal{P}U = cU \text{ and } \triangle'\,\mathcal{P}^*_{S_0}\mathcal{P}U = \mathcal{P}^*_{S_0}\mathcal{D}^2_{S_0}\mathcal{P}U = c\Lambda' U$$

where $\mathcal{D}^2_{S_0} := (-\omega_2\partial_1 + \omega_1\partial_2)^2$. To summarize, we obtain

$$c\,\Lambda' U = \mathcal{P}^*_{S_0}\mathcal{D}^2_{S_0}\mathcal{P}(U). \tag{66}$$

Generalizing the above calculation to arbitrary curves $S_0 \subset S^2$ provides the Λ-term for the reconstruction operator in ELT [151]:

$$\mathcal{R}_{\mathrm{ELT}}(g) := -\mathcal{P}^*_{S_0}\big(\mathcal{D}^2_{S_0}(g)\big). \tag{67}$$

Here, $\mathcal{D}^2_{S_0}$ is a second order differentiation in $\boldsymbol{\omega}^\perp$-plane along the tangential direction to the curve S_0, i.e.

$$\mathcal{D}^2_{S_0}(g)(\boldsymbol{\omega}, \boldsymbol{y}) := \frac{d^2}{ds^2} g(\boldsymbol{\omega}, \boldsymbol{y} + s\boldsymbol{\sigma})\Big|_{s=0} \tag{68}$$

with $\boldsymbol{\sigma}$ denoting the unit tangent to S_0 at $\boldsymbol{\omega} \in S_0$. Analogous to (66), it is shown in [56, theorem 5.1]. that if S_0 fulfills Orlov's criterion, then $\mathcal{R}_{\mathrm{ELT}}(g)$ recovers a singular pseudodifferential operator acting on U_{true} (the cited theorem does not hold for single-axis data but the result is true also for that case as shown in [151]). When Orlov's criterion is not fulfilled, then for single-axis data it turns out that the visible singularities of U_{true} correspond to singularities of $\mathcal{R}_{\mathrm{ELT}}(g)$ whenever $g = \mathcal{P}(U_{\mathrm{true}})$, see [151] and chapter ▸ Microlocal Analysis in Tomography.

Adaptation to ET

The actual reconstruction operator in ELT is given as

$$\mathcal{R}_{\mathrm{ELT}}(g) := \mathcal{P}^*_{S_0}\big(\mu - \mathcal{D}^2_{S_0}(g)\big). \tag{69}$$

This represents a local inversion method that recovers two important signal features. The pure Λ term defined in (67) emphasizes differences in data that occur at boundaries, i.e., this term picks up visible singularities. However, it does not distinguish interiors from exteriors since the derivative in these areas is typically small. The pure backprojection term (μ term) adds an averaged version of U that introduces contour to the reconstruction and allows one to distinguish objects from their surrounding.

In the implementation of ELT, the $\mathcal{D}^2_{S_0}$ operator in (69) is evaluated using a filter that is a smoothed version of the second derivative (a smoothed central second difference), and the half-width of the filter is determined by the signal-to-noise characteristics of the data. Hence, ELT has two regularization parameters, μ and the width of the derivative kernel. A method for choosing μ is given in [54] where the idea is to pick μ so that the $\mathcal{R}_{\mathrm{ELT}}$ reconstruction is as flat as possible inside a pre-specified feature. Furthermore, like in FBP, one can also convolve in the detector plane along the direction perpendicular to $\boldsymbol{\sigma}$ (see (68)), which for single-axis tilting is just averaging over slices.

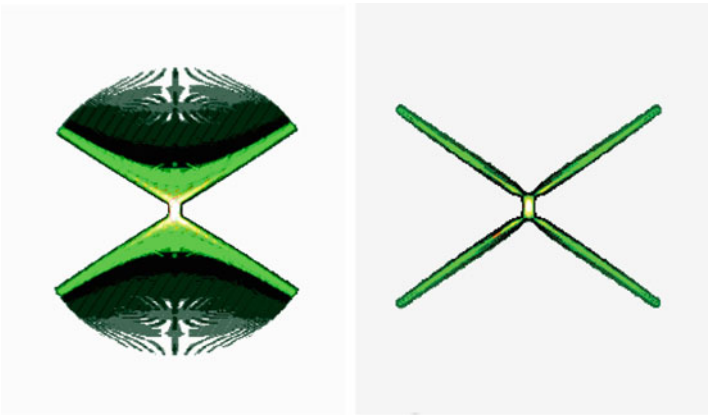


Fig. 6 A three-dimensional surface plot of the point response of FBP and ELT ($\mu = 0$) for single axis tilting (beam direction is vertical, and the tilt-axis is perpendicular to the plane) with maximum tilt-angle is 60°. Point response is more localized in ELT (*right*) than in FBP (*left*) (image from [155])

Comments and Discussion

There are several advantages in using ELT as compared to FBP/WBP. First, let us compare the convolution kernels for ELT and FBP as done in [151, Figure 1]. The ELT kernel is local, so it is zero away from the origin whereas the FBP kernel is not. In fact, the oscillations of the FBP kernel on the interval where the ELT kernel is zero are about 7 % of the maximum amplitude.

Next, ELT reconstruction operator has a more localized point response than FBP as shown in Fig. 6. The "X"-like wings at the end of the angular range in both point responses are expected in any limited angle backprojection algorithm. The point response is however less localized in the FBP case, so the artifacts spread out causing a dilution of the actual signal relative to the background. This renders the FBP more noisy than the ELT and there is a higher chance to lose a weak signal using FBP.

Generalized Ray Transform

Illuminating larger fields of view (regions that extend about 8,000 nm from the optical axis) requires using wider electron beams. Then we must account for the fact that electrons, especially those farther from the central axis, travel along helical curves. Hence, the scattering operator is better modeled by (15) instead of (16), i.e., the ray transform $\mathcal{P}$ in (57) is replaced by a generalized ray transform that integrates over curves $s \mapsto \gamma(s)$ that are solutions to (12).

The FBP approach is extended in [106] to handle such generalized ray transforms that integrate over curved electron trajectories. This paper also considers alignment, which in this new setting is far more complex, see also [3, section 4.1.2]. In a similar manner, ELT has also been extended to the above setting. This involves

characterizing visible singularities analogous to theorem 1, and deriving local inversion methods [152, 154].

Approximative Inverse

Overview

The method of approximate inverse was originally developed in [118] for solving ill-posed integral equations of first kind. It has since then been extended to a more general setting. The starting point is, as with analytic methods, the reconstruction problem with continuum data (50). The idea is to derive a reconstruction kernel that is independent of data (so it can be pre-computed), and reconstruction is obtained by applying the kernel to data. The kernel is given as a solution of an adjoint equation that involves a mollifier. The regularizing property of approximate inverse comes from the choice of mollifier and robustness of the kernel against noise in data.

The approximative inverse method is by now a well-established approach for regularization with applications to a wide range of inverse problems, see the survey by [170]. Recent developments include extending the framework to a Banach space setting and to handle certain nonlinearities in the forward problem [171, part V]. See also [115] for an interesting unification concept for general regularization methods based on approximate inverse. The application to ET is given in [101, 102].

A Priori Information

As the name suggests, the original formulation of approximate inverse seeks provide an approximative inverse solution to the inverse problem. In this setting, the prior information is the same as for backprojection-based methods (section "Backprojection-Based Methods"). The approximate inverse method is in some sense the natural generalization of the FBP/WBP method to the setting where the forward operator is a general affine operator, like in (57).

Data Pre-processing

Computationally feasible implementations require an affine forward operator, which in particular includes operators of the type in (57).

Reconstruction Operator

We describe the reconstruction operator for linear operators in the Hilbert space setting and refer to [170] for its extensions to more general settings.

Consider a linear forward operator $\mathcal{T}: \mathscr{X} \to \mathscr{H}$ between Hilbert spaces $\mathscr{X}$ and $\mathscr{H}$, each with inner products $\langle \cdot, \cdot \rangle_{\mathscr{X}}$ and $\langle \cdot, \cdot \rangle_{\mathscr{H}}$, respectively. Next, let $E_\gamma: \mathbb{R}^3 \times \mathbb{R}^3 \to \mathbb{C}$ be a mollifier for $\mathscr{X}$. The formal definition in the Hilbert space setting is given on [170, p. 26], but for our purposes it is enough to think of E_γ as a smoothing operator that converges to a δ-distribution as $\gamma \to 0$:

$$\left\langle E_\gamma(\boldsymbol{x}, \cdot), U \right\rangle_{\mathscr{X}} \to U(\boldsymbol{x}) \quad \text{as } \gamma \to 0 \text{ for all } U \in \mathscr{X}.$$

Fix a point $\boldsymbol{x}$ and a mollifier and define the reconstruction kernel $\phi_{\boldsymbol{x}}^{\gamma} \in \mathscr{H}$ as the solution to the adjoint equation

$$\mathcal{T}^*\big(\phi_{\boldsymbol{x}}^{\gamma}\big) = E_{\gamma}(\boldsymbol{x}, \cdot). \tag{70}$$

Note that the above calculation does not depend on data. Moreover, if the above equation lacks a solution, we consider the minimal norm solution, i.e., $\phi_{\boldsymbol{x}}^{\gamma}$ is the element in $\mathscr{X}$ that minimizes the $\mathscr{X}$-norm of $\phi \mapsto \mathcal{T}^*\big(\phi\big) - E_{\gamma}(\boldsymbol{x}, \cdot)$. The reconstruction operator $\mathcal{R}_{\mathrm{AI}}: \mathscr{H} \to \mathscr{X}$ is now defined as

$$\mathcal{R}_{\mathrm{AI}}(g)(\boldsymbol{x}) := \big\langle g, \phi_{\boldsymbol{x}}^{\gamma} \big\rangle_{\mathscr{H}} \quad \text{for } g \in \mathscr{H}. \tag{71}$$

The rationale for (71) is that it provides a mollified solution to our inverse problem. To see this, let $g = \mathcal{T}(U_{\mathrm{true}})$, so

$$\begin{aligned} \mathcal{R}_{\mathrm{AI}}(g)(\boldsymbol{x}) &= \big\langle g, \phi_{\boldsymbol{x}}^{\gamma} \big\rangle_{\mathscr{H}} = \big\langle \mathcal{T}(U_{\mathrm{true}}), \phi_{\boldsymbol{x}}^{\gamma} \big\rangle_{\mathscr{H}} \\ &= \Big\langle U_{\mathrm{true}}, \mathcal{T}^*\big(\phi_{\boldsymbol{x}}^{\gamma}\big) \Big\rangle_{\mathscr{X}} = \big\langle U_{\mathrm{true}}, E_{\gamma}(\boldsymbol{x}, \cdot) \big\rangle_{\mathscr{X}}. \end{aligned}$$

Next, calculating the reconstruction kernel requires solving the adjoint equation (70), which can be done in three different ways: (1) By means of an inversion formula for $\mathcal{T}$, (2) through a singular value decomposition, and (3) to use a projection method that recasts the adjoint equation into the finite dimensional setting where it can be solved using numerical methods.

One issue is that the adjoint equation (70) needs to be solved for each evaluation point $\boldsymbol{x}$. This is clearly unfeasible for 3D tomographic reconstruction problems. If the inverse problem is translation invariant, then it is natural to use convolution type mollifiers that only depend on the difference of the argument: $E_{\gamma}(\boldsymbol{x}, \boldsymbol{y}) = e_{\gamma}(\boldsymbol{x} - \boldsymbol{y})$. Given such mollifiers, the structure of the adjoint equation (70) does not depend on the point $\boldsymbol{x}$:

$$\mathcal{T}^*\big(\phi^{\gamma}\big) = e_{\gamma}. \tag{72}$$

Hence, the reconstruction kernel ϕ^{γ} is independent of $\boldsymbol{x}$ and the reconstruction operator yields $\mathcal{R}_{\mathrm{AI}}(g) = U_{\mathrm{true}} * e_{\gamma}$ whenever $g = \mathcal{T}(U_{\mathrm{true}})$.

Adaptation to ET

The adaptation to ET lies in the computation of the reconstruction kernel which in turn depends on the mollifier and the forward operator. The choice of the mollifier should match the a priori information about U_{true} and choice of γ (regularization parameter) depends on the noise level in data. Since ET is a translation invariant inverse problem, the reconstruction kernel can be computed given these choices by solving (72).

In ET we consider $\mathcal{T}$ as the linear part of (57), i.e., if $\mathcal{A}$ convolves with PSF in (57) then (72) becomes $(\mathcal{P}_{S_0}^* \circ \mathcal{A}^*)(\phi^{\gamma}) = e_{\gamma}$. We can now first compute v^{γ} by solving $\mathcal{P}_{S_0}^* \circ v^{\gamma} = e_{\gamma}$, and then compute the reconstruction kernel ϕ^{γ} from

$\mathcal{A}^*(\phi^\gamma) = v^\gamma$. A common mollifier is to take $e_\gamma(\boldsymbol{x}) := \gamma^{-1}\tau(\boldsymbol{x}/\gamma)$ where τ is a scaling function. $\mathrm{PSF}(\boldsymbol{\omega}, \cdot)$ has a closed form expression in Fourier space, which provides a closed form expression for the reconstruction kernel in Fourier space for parallel beam geometries relevant for ET. Hence, in ET the convolution with ϕ^γ that defines the reconstruction operator is preferably computed in Fourier space, see [101, 102] for details. A special case is when the $\mathrm{PSF}(\boldsymbol{\omega}, \cdot) = \delta_{\boldsymbol{\omega}^\perp}$ in (57), i.e., we consider the inversion of the ray transform. The corresponding adjoint equation reads as $\mathcal{P}^*_{S_0}(\phi^\gamma) = e_\gamma$, so $\mathcal{R}_{\mathrm{AI}}(g) = \mathcal{R}_{\mathrm{FBP}}(g)$ where h in (60) given by ϕ^γ.

Comments and Discussion

The benefit of approximative inverse, as compared to iterative (section "Iterative Methods with Early Stopping") and variational (section "Variational Methods") methods, lies in the combination of computational efficiency and flexibility regarding the type of forward operators it can handle, and to some extent, the type of a priori information it can encode. It is in fact easy, at a first glance, to underestimate the diversity of the a priori information that the approximate inverse method can encode.

First, it is possible to work with a mollifier that is not rotation invariant [116, section 3]. This can be beneficial for ET since resolution is anisotropic due to the limited angle problem. Another possibility is given in [117] where approximative inverse is used for combined reconstruction and feature detection. The feature detection is here represented by a linear operator $\mathcal{L}\colon \mathscr{X} \to \mathscr{Y}$ ($\mathscr{Y}$ is the Hilbert space of features). To compute an approximate inverse to $\mathcal{L}(U_{\mathrm{true}})$ requires us to modify the adjoint equation (70) that provides the reconstruction kernel. In this new setting, it becomes $\mathcal{T}^*\big(\phi^\gamma_{\boldsymbol{x}}\big) = \mathcal{L}\big(E_\gamma(\boldsymbol{x}, \cdot)\big)$. If $\mathcal{L}$ is translational invariant, the adjoint equation is again independent of the evaluation point $\boldsymbol{x}$ and it can be written as $\mathcal{T}^*\big(\phi^\gamma\big) = \mathcal{L}\big(e_\gamma\big)$. The case when $\mathcal{L}$ is an edge detector is treated in [117]. One can also select $\mathcal{T} = \Lambda$ and thereby perform Λ-tomography (section on p. 994), so approximate inverse extends ELT to the setting where one also includes deconvolution of the PSF.

A drawback with the approximative inverse method is that it is difficult (if not impossible) to devise schemes for selecting the filter γ that takes into account the specific stochasticity of the data.

Iterative Methods with Early Stopping

Iterative methods are a broad class of well-studied methods, see [11, 50, 77] for good surveys of both theory and implementation. The idea is to construct an iterative scheme in $\mathscr{X}$ that, in the limit, converges to a solution of (55), i.e., one of possibly infinitely many ML/least-squares solutions of (53).

Overview

There are three classes of iterative methods, *conjugate gradient*, *Maximum-Likelihood Expectation Maximization (ML-EM)*, and *iterative algebraic* methods. These have over the years been extended in various ways, e.g., to handle nonlinear

forward operators, to account for different types of a priori information (like positivity), and to handle more complex noise models for data. Surveys of these, and other developments, are provided in [25, 81] for iterative algebraic techniques, in [92] and chapter ▸ Iterative Solution Methods for conjugate gradient type of methods, and in chapter EM Algorithms for ML-EM type of methods.

In the context of ET, only iterative algebraic methods are used. Hence, we focus on this category following the excellent exposition in [136, section 5.3], see also chapter ▸ Tomography. The main idea is to split the inverse problem in (53) into a finite number of sub-problems. A cyclic iteration over these sub-problems is required to generate the next iterate. The specific choice of splitting leads to different iterative algebraic methods, such as the Algebraic Reconstruction Technique (ART) and the Simultaneous Iterative Reconstruction Technique (SIRT). The common theme is however to generate an iterative sequence that, in the limit, converges to a least-squares solution of (53).

Data Pre-processing

Same as for the approximate inverse method (section "Approximative Inverse").

A Priori Information

Iterative methods with early stopping share a common a priori assumption, namely that the iterative scheme is *semi-convergent*. This means that initial iterates recover low-frequency components of the signal in (53). Now, assume noise from data mainly influences the high-frequency components of the signal. Then, an ill-posed problem can be regularized by stopping the iterates before too much of the data noise impairs the reconstruction, i.e., the number of iterates becomes a regularization parameter. Iterates are also often smoothed for further regularization.

One can also account for additional a priori information. A common case is to require that U_{true} is a positive function, or more generally, belongs to a convex set. Such constraints can be enforced by projecting iterates onto the said convex set, albeit at a significant computational cost.

Reconstruction Operator

Consider (53) and split the data (tilt-series) $\boldsymbol{g} \in \mathbb{R}^m$ into N subsets:

$$\boldsymbol{g} = \left(\boldsymbol{g}_1, \ldots, \boldsymbol{g}_N\right) \quad \text{with } \boldsymbol{g}_j \in \mathbb{R}^{m_j} \text{ and } m = m_1 + \cdots + m_N. \tag{73}$$

Next, let $\tau_j : \mathbb{R}^m \to \mathbb{R}^{m_j}$ be the projection onto the j:th data component, so $\tau_j(\boldsymbol{g}) := \boldsymbol{g}_j$, and introduce the partial forward operator $\mathcal{T}_j : \mathcal{X} \to \mathbb{R}^{m_j}$ as

$$\mathcal{T}_j(U) := (\tau_j \circ \mathcal{T})(U) \quad \text{for } j = 1, \ldots, N.$$

Then (53) splits into the following set of N sub-problems:

$$\boldsymbol{g}_j := \mathcal{T}_j(U_{\text{true}}) + \boldsymbol{g}_j^{\text{noise}} \quad \text{where } \boldsymbol{g}_j^{\text{noise}} := \tau_j(\boldsymbol{g}^{\text{noise}}). \tag{74}$$

The k:th iterate in an iterative algebraic method is obtained by performing an inner N-step iteration:

$$\begin{cases} U_k^0 := U_{k-1} \\ U_k^j := U_k^{j-1} + \mu\big(\boldsymbol{\mathcal{T}}_j^* \circ \boldsymbol{C}_j^{-1}\big)\big(\boldsymbol{g}_j - \boldsymbol{\mathcal{T}}_j(U_k^{j-1})\big), \quad j = 1, \ldots, N \\ U_k := U_{k-1}^N. \end{cases} \tag{75}$$

Here, $\boldsymbol{C}_j : \mathbb{R}^{m_j} \to \mathbb{R}^{m_j}$ is a fixed linear positive definite operator, and a typical choice for linear $\boldsymbol{\mathcal{T}}_j$ is $\boldsymbol{C}_j := \boldsymbol{\mathcal{T}}_j \circ \boldsymbol{\mathcal{T}}_j^*$. The parameter $\mu > 0$ is a relaxation parameter that is needed for noisy problems, like the one in ET, to reduce the impact of noise in data and speed up convergence.

Remark 22. There is a geometric interpretation of the iterates in (75). Each can be seen as the result of performing N consecutive projections in reconstruction space $\mathscr{X}$. More precisely, the j-iteration in (75) can be written as $U_k^j := \pi_j^\mu(U_k^{j-1})$ where $\pi_j^\mu(U) := (1-\mu)U + \mu\,\pi_j(U)$ with $\pi_j : \mathscr{X} \to \mathscr{X}$ denoting the projection operator

$$\pi_j(U) := U + \big(\boldsymbol{\mathcal{T}}_j^* \circ \boldsymbol{C}_j^{-1}\big)\big(\boldsymbol{g}_j - \boldsymbol{\mathcal{T}}_j(U)\big) \quad \text{for } j = 1, \ldots, N.$$

Note that when $\boldsymbol{C}_j := \boldsymbol{\mathcal{T}}_j \circ \boldsymbol{\mathcal{T}}_j^*$, then $\pi_j(U)$ is merely the projection of the residual $\boldsymbol{g}_j - \boldsymbol{\mathcal{T}}_j(U)$ onto the solution space of $\boldsymbol{\mathcal{T}}_j(U) - \boldsymbol{g}_j = \boldsymbol{0}$.

The specific choice of splitting and choice of $\boldsymbol{C}_j$ determines the type of iterative algebraic method. The three most common ones are listed below:

Algebraic Reconstruction Technique (ART): This is when data (73) is split up so that $\boldsymbol{g}_j$ is a scalar corresponding to a single data point, i.e., $N = m$, and $\boldsymbol{C}_j = \boldsymbol{\mathcal{T}}_j \circ \boldsymbol{\mathcal{T}}_j^*$. The iterative sequence in (75) is now easier to understand when expressed in the full discretized setting. The (fully discretized) linear forward operator is given by multiplication with a $(m \times n)$-matrix whose rows are denoted by $\boldsymbol{a}_j \in \mathbb{R}^n$ for $j = 1, \ldots, m$. The corresponding sub-problems (74) are now given as

$$g_j = \boldsymbol{a}_j \cdot \boldsymbol{U}_{\text{true}} + g_j^{\text{noise}} \quad \text{for } j = 1, \ldots, m,$$

and the projection calculation in (75) is expressible as

$$\boldsymbol{U}_k^j := \boldsymbol{U}_k^{j-1} + \mu \frac{1}{\|\boldsymbol{a}_j\|^2}\big(g_j - \boldsymbol{a}_j \cdot \boldsymbol{U}_k^{j-1}\big) \cdot \boldsymbol{a}_j \quad \text{for } j = 1, \ldots, m.$$

The ART, first introduced in [70], is perhaps the first iterative algebraic technique used for tomographic reconstruction. Later it was recognized it as a special case of Kaczmarz's method.

Simultaneous Iterative Reconstruction Technique (SIRT): Compared to ART, SIRT is the other extreme in the sense that data is not slit, i.e., $N = 1$. Hence, SIRT reconstruction is given as

$$U_{k+1} := U_k + \mu \mathcal{T}^* \big(\boldsymbol{g} - \mathcal{T}(U_k) \big). \tag{76}$$

Thus, each iterate corresponds to a full update of the signal that involves a complete sweep through all the data points.

The SIRT was introduced to the scientific community in the context of tomography by [65] as an alternative to ART. It was later discovered that the SIRT is equivalent to Landweber iteration.

Simultaneous ART (SART): SART, also called block iterative ART, was first introduced to the imaging community by [5]. It is a compromise between ART and SIRT in the sense that N equals the number of views. For ET this means that N is the number micrographs in the tilt-series, i.e., $N = m_{\text{tlt}}$ and $\boldsymbol{g}_j \in \mathbb{R}^{n_{\text{det}}}$ corresponds to the j:th micrograph.

The regularizing property of iterative algebraic techniques is far from resolved in the general setting. Most of the literature focuses on proving convergence properties, as reviewed in [25]. Such results are of less interest for ill-posed problems since iterates are stopped long before convergence (assuming the iterates do converge to some point). A fast convergence is nonetheless desirable since this would require fewer iterates before a "good" reconstruction is obtained. In general, the convergence of iterative algebraic methods depends on the splitting. For tomographic data, a good strategy is to do a splitting so that the directions are as orthogonal as possible to the previous ones. Curiously enough, a random choice of directions is almost as good as the optimal choice. This was also recently proved mathematically in [26].

Adaptation to ET

The adaptation of iterative algebraic methods to ET lies in the choice of forward operator and its adjoint, and in how to select the number of iterates $k_{\max}$ and the relaxation parameter μ.

In ET it is rather natural to split the tilt-series into micrographs (2D TEM images), so current implementations of iterative algebraic methods are based on SART rather than ART (even though this is not explicitly mentioned). Modern implementations also include possibility for more involved projection schemes, like component and block component averaging.

As of writing, in the context of ET there is no theory that provides a practically useful criteria for how to choose the number of iterates $k_{\max}$. This is perhaps not that big of an issue since one can always run a couple of extra iterates. If one saves intermediate iterates, each such iterate corresponds to a specific choice of $k_{\max}$. Concerning μ, an analysis of how its choice influences the qualitative behavior of the iterates must exploit the specific structure of the forward operator. For the 2D ray transform one can show that a strong under-relaxation, i.e., small μ results in iterations that first determine the smooth parts of the signal, while the higher resolution details appear later [136, pp. 113–115]. It is reasonable to expect that the

same holds for a forward operator of the type (57), so the common suggestion in ET is $\mu = 0.01$ (recommended range between 0 and 0.1).

Yet another aspect is related to how the reconstruction space $\mathscr{X}$ is discretized in an implementation. The straight-forward approach is to simply evaluate the function at each voxel center. In [64, 109] one claims that a discretization based on Kaiser–Bessel window functions (blobs) improves the robustness against noise.

Finally, we consider the issue of a priori information. A variety of a priori information relevant for ET can be incorporated into iterative algebraic methods. One can beforehand determine regions where the signal has to attain a given value (say zero), one can set upper and lower values for the signal, or only update parts of the signal. Such Projection Onto Convex Sets (POCS)-type of constraints, that can be formulated as the signal belonging to a convex set, can be enforced by projecting iterates. POCS was introduced to the electron microscopy community in [28], but convincing evidence of benefits is lacking in ET [145, p. 315]. Iterative algebraic methods can also be used to solve variational regularization methods (section "Variational Methods") with quadratic energy functionals. A final development is the so-called discrete ART where one discretizes the range of the allowed reconstructions. This is obviously suitable for specimens that consist of only a few different materials (gray levels), like in material sciences, and results are encouraging [13, 72].

Comments and Discussion

Iterative algebraic methods offer flexibility regarding the type of forward operators it can handle, and to some extent, the type of a priori information it can encode. Compared to approximate inverse (section "Approximative Inverse"), the flexibility is greater, but so is the computational burden.

SIRT is considered to give smoother reconstructions, but converges more slowly. Still, it is now part of most software packages for ET in both life and material sciences. Both SART and SIRT perform better than FBP/WBP [120], even though the difference is not that dramatic. There are in fact few situations where the significantly longer run-time associated with SART/SIRT is worth the effort. The run-time of SART/SIRT is however becoming less of an issue thanks to recent parallel implementations that make use of hardware acceleration. These allow reconstructions within a reasonable time of regions as large as $2048 \times 2048 \times 512$ voxels from tilt-series of 60–80 micrographs, each with a size of 2048×2048 pixels [177]. We finally mention [72] where SIRT, total variation regularization (section on p. 1008), and the discrete algebraic reconstruction technique are compared for material sciences application of ET.

Variational Methods

Variational methods provide a flexible framework for regularization of (53). Here, the reconstruction is given as a solution to an optimization problem:

$$\min_{U \in \mathscr{X}} \ \mu \mathcal{S}(U) + \mathcal{D}\Big(\boldsymbol{\mathcal{T}}(U), \boldsymbol{g}\Big) \quad \text{for } \boldsymbol{g} \in \mathbb{R}^m \text{ given as in (53).} \tag{77}$$

In the above, $\boldsymbol{\mathcal{T}}: \mathscr{X} \to \mathbb{R}^m$ is the sampled forward operator, which for ET is given by (52), $\mathcal{S}: \mathscr{X} \to \mathbb{R}_+$ is the *regularization functional* that enforces uniqueness and stability by incorporating a priori knowledge about $U_{\text{true}} \in \mathscr{X}$, $\mathcal{D}: \mathbb{R}^m \to \mathbb{R}_+$ is the *data discrepancy functional* that quantifies the goodness-of-fit against measured data, and $\mu > 0$ is a *regularization parameter* quantifying the compromise between accuracy and stability. Its choice should depend on the model for the noise, as well as on estimates of the size of the error. Formally, without knowledge of the latter, it is not possible to guarantee a result with low error, neither absolutely nor as a percentage [10].

If (77) is to be a regularization of (53), then solving (77) must be well-posed. Such issues form a central part of the general theory, and current state of the art is nicely surveyed by [169] in the context of imaging.

In the fully discrete setting (54), there is also a nice connection to statistical regularization. If $\mathcal{D}$ is chosen as an affine transformation of the negative log-likelihood of the random variable modeling $\boldsymbol{g}$, then (77) corresponds to a maximum a posteriori estimate where the prior probability distribution is given as a Gibbs measure with an energy $\mathcal{S}$ [91, subsection 3.3.2]. Extending this analogy to (53) is more elaborate since it requires a notion of Gibbs measure on the Hilbert/Banach space $\mathscr{X}$.

Data Pre-processing

Same as for the approximate inverse method (section "Approximative Inverse").

Entropy Regularization

Overview

Entropy regularization refers to the case when the regularization functional $\mathcal{S}$ in (77) is given by an entropy type of functional. There are plenty of examples, mainly from astronomy, where entropy type of functionals are used for regularization.

Existence, uniqueness, stability, convergence, and convergence rate for the maximum entropy method are studied in [47] when the forward operator is a Fredholm integral operator, but most results seem to be valid for an arbitrary bounded linear operator from $\mathscr{L}^1$ to a Hilbert space. A similar setting is dealt with in [49], which obtains stability, convergence, and convergence rate results. A wider class of regularization functionals is considered in [159], where convergence rates and error estimates are proved when the forward operator is a compact linear operator between Banach spaces, see also [137] for the case when the forward operator is given by convolution. Likewise, [82] establishes existence and regularity of solutions for convolution type of forward operators and a wide range of entropy type of regularization functionals.

The application to ET is in [172] with a mathematical analysis in the language of regularization theory given in [165].

A Priori Information

It is clear that entropy type of functionals do penalize complexity in some sense, thereby acting as a regularization. Entropy type of functionals also measure information content [44], so many attempts to rationalize the usage of entropy for regularization are based on this principle. This motivation is artificial unless there is a natural probability distribution on $\mathscr{X}$. This leads us to the framework of statistical regularization in which entropy is applied to the set of probability measures on $\mathscr{X}$ [119, Method 3], but that is an entirely different approach for reconstruction.

Another motivation is provided by [38, 174] that considers reconstruction methods (called selection rules) for linear inverse problems with exact data. One starts off by postulating a set of axioms (consistency, distinctness, continuity, locality, and composition consistency) that a reconstruction method should satisfy. Next, it is proved that the least squares method is the only reconstruction method consistent with these axioms when the signal to be reconstructed is a real-valued function, having both negative and positive values, and entropy regularization is the only reconstruction method consistent when the signal to be reconstructed is positive.

Reconstruction Operator

The case we consider here makes use of the relative entropy, so we are given a fixed prior $\rho \in \mathscr{X}$ and the regularization functional is

$$\mathcal{S}_\rho(U) := \int_\Omega \left(U(\boldsymbol{x}) \ln \left(\frac{U(\boldsymbol{x})}{\rho(\boldsymbol{x})} \right) - U(\boldsymbol{x}) + \rho(\boldsymbol{x}) \right) \mathrm{d}\boldsymbol{x} \quad \text{for } U \in \mathscr{X}. \tag{78}$$

Then, $U \mapsto \mathcal{S}_\rho(U)$ is convex and $\mathcal{S}_\rho(U) \geq 0$. Note also that when $U, \rho \in \mathscr{X}$ represent probability measures, then $\mathcal{S}_\rho(U) = 0$ if and only if $U = \rho$ almost everywhere and $-\mathcal{S}_\rho(U)$ equals the Kullback–Leibler divergence.

Adaptation to ET

In [172] one regularizes the inverse problem in ET by (77) with (78) and a data discrepancy given by a Mahalanobis distance (section "Notion of Solution"). The forward operators are as in (57) and the regularization parameter is chosen according to the Morozov principle, so a user must supply an estimate of the data error.

The Morozov principle for choosing the regularization parameter is not usable for highly noisy data since it requires unreasonably accurate estimates of the data error. Furthermore, a higher reconstruction resolution requires one to estimate the nuisance parameters. Thus, the approach in [172] is extended in [165], which uses an iterated regularization scheme (not to be confused with iterative regularization in section "Iterative Methods with Early Stopping") that generates a sequence of regularization problems. Within this sequence, the estimate of the data error is "updated." The nuisance parameters $\boldsymbol{\omega} \mapsto C(\boldsymbol{\omega})$ and C_0 are recovered by *intertwining* the iterates that update the signal $U \in \mathscr{X}$ with least-squares iterates that update the nuisance parameters and locality in tomographic data is accounted for by constant continuation (section on p. 985). To formulate the precise scheme, for given data $\boldsymbol{g} \in \mathbb{R}^m$ and nuisance parameters $c \in \mathscr{V}$ (the space of nuisance

parameters), we introduce

$$\epsilon_{\min}(\boldsymbol{g}, c) := \inf_{U \in \mathscr{X}} \left\| \mathcal{T}(U, c) - \boldsymbol{g} \right\|_{\mathbb{R}^m}^2.$$

Hence, $\epsilon_{\min}(\boldsymbol{g}, c)$ is the smallest possible data error (it is zero for exact data). Also, define $\mathcal{A}_b$ as a smoothing operator whose degree of smoothing is regulated by a parameter $b > 0$, it can, e.g., be a low-pass filter where b is the cut-off threshold in the Fourier space. Finally, $0 \leq \delta \leq 1$ is the regularization parameter that governs the updating of the estimate of the data error. Then, for a user provided choice of regularization parameters $\delta, b > 0$, the iterates $(U_j, c_j) \in \mathscr{X} \times \mathscr{V}$ in [165] are defined recursively as

$$c_j := \operatorname{argmin}_{c \in \mathscr{V}} \left\| \mathcal{T}(U_{j-1}, c) - \boldsymbol{g} \right\|_{\mathbb{R}^m} \tag{79}$$

$$\rho_j := \begin{cases} \mathcal{A}_b(U_{j-1}) & \text{if prior is to be updated,} \\ \rho & \text{if prior is not updated,} \end{cases} \tag{80}$$

$$\epsilon_j := \epsilon_{\min}(\boldsymbol{g}, c_j) + \delta \left(\left\| \mathcal{T}(\rho_j, c_j) - \boldsymbol{g} \right\|_{\mathscr{H}} - \epsilon_{\min}(\boldsymbol{g}, c_j) \right) \tag{81}$$

$$U_j := \begin{cases} \operatorname{argmin}_{U \in \mathscr{X}} \mathcal{S}_{\rho_j}(U) \\ \left\| \mathcal{T}(U, c_j) - \boldsymbol{g} \right\|_{\mathbb{R}^m}^2 \leq \epsilon_j. \end{cases} \tag{82}$$

Comments and Discussion

Entropy type of regularization performs fairly well, especially on cryo-fixated in vitro specimens that contain isolated particles in aqueous environment. The idea of updating the estimate of the data error through (81) by choosing $0 < \delta < 1$ is much more robust than directly specifying an estimate for it in (82). On the other hand, one has to be careful in updating of the prior (80). If the prior is updated, reconstructions appear de-noised and de-speckled, but there is a significant risk of creating structures that tend to grow as iterates proceed and therefore are erroneously interpreted as true structures. If the prior is to be updated, it is highly recommended that one uses an edge preserving filter as $\mathcal{A}_b$. Finally, the nuisance parameters can equally well be estimated off-line by a least-squares approach, removing the need for (79).

TV Type of Regularization

Overview

We use the term "TV type" of regularization for variational methods in which $\mathcal{S}$ in (77) is chosen as

$$\mathcal{S}(U) := \left(\int_{\Omega_0} \left| \nabla U(\boldsymbol{x}) \right|^p \mathrm{d}\boldsymbol{x} \right)^{1/p}. \tag{83}$$

Standard TV regularization is when $p = 1$, see chapters ▸ Total Variation in Imaging and ▸ Numerical Methods and Applications in Total Variation Image Restoration for surveys of its role in imaging. The emphasis here is on its usage in ET.

A Priori Information

Standard TV regularization is based on the a priori assumption that the signal to be recovered has a sparse gradient, e.g., like a step function.

Reconstruction Operator

The reconstruction operator is given as the solution of (77) with $\mathcal{S}$ as in (83) and $\mathcal{D}$ a suitable data discrepancy functional (section "Notion of Solution"). If the forward operator is affine, and the data discrepancy is strictly convex, then (77) has a unique solution.

Adaptation to ET

The first usage of TV regularization in ET was [1]. The approach is specifically tailored towards single axis tilting data (section "Examples Relevant for ET") since one performs a series of 2D reconstruction in slices orthogonal to the tilt-axis. The regularizing functional is an anisotropic variant of TV that accounts for the anisotropy in resolution due to limited angle data (section "Incomplete Data, Uniqueness and Stability"). Standard TV regularization has also been applied in [71] on STEM data of specimens from material sciences (nano-particles). Results are compared against SIRT and TV regularization performs significantly better.

Both publications above that deal with TV regularization for ET have a number of weaknesses. First, no guidance is offered on how to select the regularization parameter. Next, both publications unnecessarily assume a forward operator given as the ray transform, so ET data needs to be pre-processed in the same way as for backprojection-based methods (section "Backprojection-Based Methods"). This is not an issue when an amplitude contrast model (section "Forward Operator for Amplitude Contrast Only") is good enough for modeling micrograph contrast, like for incoherent STEM imaging in material sciences, but it is inadequate for ET based on HRTEM data. Finally, the data discrepancy functional $\mathcal{D}$ is the 2-norm, corresponding to the assumption of additive Gaussian noise whereas the actual noise is more complex. This is however probably not that big of an issue (paragraph on p. 976).

The first two points raised above are in fact addressed in [164] that considers variational regularization of (53) with an affine forward operator of the form (57). The data discrepancy functional $\mathcal{D}$ in (77) is given by the 2-norm and the regularization functional $\mathcal{S}$ is of the form

$$\mathcal{S}(U) := \lambda \left(\int_{\Omega_0} \left| \nabla U(\boldsymbol{x}) \right|^p \mathrm{d}\boldsymbol{x} \right)^{1/p} + \mu \left(\int_{\Omega_0} \left| U(\boldsymbol{x}) \right|^q \mathrm{d}\boldsymbol{x} \right)^{1/q}. \tag{84}$$

Hence, choosing $p = 1$ and $\mu = 0$ gives the usual TV regularization. The implementation works with a forward operator that models both amplitude and phase contrast for any parallel beam geometry. Nuisance parameters $\boldsymbol{\omega} \mapsto C(\boldsymbol{\omega})$ and C_0 are estimated by least-squares from tilt-series data and locality in tomographic data is accounted for by constant continuation (section on p. 985). The main scientific contribution of [164] is however a method for choosing the regularization parameter that is specifically designed for highly noisy ET data.

Comments and Discussion

Variational methods, like TV regularization, often perform very well in reducing speckle and noise if the choice of regularization functional correctly encodes some of the a priori regularity properties of U_{true}. An issue is however their computational feasibility, and this is especially so for TV regularization in 3D imaging. The non-differentiability of the regularization functional in TV makes it difficult to directly use efficient gradient-based methods for solving (77). There are several iterative approaches for this purpose, see chapter ▸ Numerical Methods and Applications in Total Variation Image Restoration for further details. This issue of computational feasibility is also related to choosing the regularization parameter. In contrast to iterative methods (section "Iterative Methods with Early Stopping"), for variational approaches each choice of regularization parameter requires a new iterative sequence. Hence, a critical part of computational feasibility is the ability to choose the regularization parameter.

Another issue with TV regularization relates to the suitability of using the (83) with $p = 1$. This results in reconstructions that often have too narrow dynamic range and suffer from stair-casing. This might not be an issue for automated image analysis tasks, like segmentation, but biologists consider such reconstructions to have an "unnatural" appearance. As an alternative to standard TV regularization, one could consider using other TV-like energy (semi) norms, e.g., empirical tests show that using $1.1 \leq p \leq 1.5$ in (84) gives reconstructions that compare favorably to TV (the case $p = 1$) regarding noise/speckle reduction, while gray-scale variations are recovered better. One can also consider Besov norms that are discretization-invariant and better at recovering smoothly varying parts of the signal [105]. An interesting approach is to consider the norm parameter p as a nuisance parameters and try to estimate it from data, like in [182] for the Besov norm. Another is to combine TV regularization with Bregman iterates and/or use higher-order TV methods to better recover smoothly varying parts of the signal and reduce stair-casing [14]. An example, [193] considers a variant of Dirichlet regularization that makes use of geometric information. This corresponds to choosing the regularization functional $\mathcal{S}$ in (77) as

$$\mathcal{S}(U) := \int_{\Omega_0} u\big(\sigma(U)(\boldsymbol{x}), \kappa(U)(\boldsymbol{x})\big) \big| \nabla U(\boldsymbol{x}) \big|^2 \, \mathrm{d}\boldsymbol{x}$$

where $\sigma(U)(\boldsymbol{x})$ denotes the mean Gaussian curvature of the level-set surface of U at $\boldsymbol{x}$, and $u\colon \mathbb{R}^2 \to \mathbb{R}_+$ is a user specified function. The optimization problem is by

solving an L^2-gradient flow using calculus of variations. The method is compared against SART, SIRT, and WBP on a tilt-series of a cryo-fixated in vitro specimen with isolated HIV virions in aqueous buffer.

Finally, one may also consider using a norm in a variable Lebesgue space (Theorem 2.15 and 2.17 in [36]). Already in [18] it was suggested that in image reconstruction, a smoother image could be obtained by an interpolation technique that uses a variable exponent that decreases monotonically from two to one as ∇U increases, see comments in pp. 7–8 in [36]. This however leads to a non-convex functional that is computationally difficult to handle. One approach is then use Λ-tomography to determine the variable exponent. More precisely, consider the regularization functional

$$\mathcal{S}(U) := \int_{\Omega_0} \left| \nabla U(\boldsymbol{x}) \right|^{p(|\nabla \rho(\boldsymbol{x})|)} \mathrm{d}\boldsymbol{x} \tag{85}$$

where $p\colon \mathbb{R}_+ \to [1, 2]$ is monotonically decreasing, e.g., $p(t) = 2/(1 + 2t)$, and ρ is pre-computed from data. If the singularities of ρ are at the same location as U_{true}, then using (85) instead of (83) yields much smoother reconstructions in regions of moderate gradient and thus prevents stair-casing [33]. One can use ELT to calculate ρ and use microlocal analysis to characterize the singularities of ρ that coincide with those of U_{true} (section on p. 994). Albeit convex, the downside of using (85) is that it is not 1-homogenous, so multiplicative scaling might change the values of the regularization functional.

Sparsity Promoting Regularization

Overview

As we have seen, sampling theory is at the heart of many image reconstruction algorithms for tomography, like FBP/WBP (section "Backprojection-Based Methods"). These methods make the assumption that noise is predominantly a high frequency phenomena, whereas relevant image features occupy low frequencies of the signal. Most signals are however sparse (compressible) is some suitable representation. It turns out that this a priori knowledge, if properly utilized, allows one to recover the signal from relatively few and/or highly noisy measurements.

A straightforward way to take advantage of sparsity often leads to solving (77) with an 0-norm as regularization functional, which unfortunately is computationally unfeasible. Within geophysics and other scientific/engineering disciplines, it has been known since the late 1970s that one can use the computationally feasible 1-norm instead of the 0-norm is certain cases. This observation was put into a formal mathematical setting in 2004 with the advent of a new sampling theory, compressed sensing, which puts precise conditions on a linear inverse problem for when the 1-norm on $U \mapsto \rho(U)$ in (77) gives the sparsest solution [27]. The field has since then exploded with several remarkable and far-reaching results, see chapter ▸ Compressive Sensing and the recent book [59] for an up-to-date survey. Below we focus on the role of sparsity in ET.

A Priori Information

Let us start by formalizing the notion of sparsity. We say that $\rho: \mathscr{X} \to \mathscr{Y}$ is a sparsifying map for $U_{\text{true}} \in \mathscr{X}$ whenever $\rho(U_{\text{true}})$ is "well approximated" by a low-dimensional subspace of $\mathscr{Y}$ ($\mathscr{Y}$ is here some suitable vector space). A common example is when there is a fixed set $\{\phi_i\}_i \subset \mathscr{X}$ (dictionary) such that $U \sim \sum_i \alpha_i \phi_i$ with most $\alpha_i \approx 0$. In such case, $\mathscr{Y}$ is the space of sequences and $\rho(U) := \{\alpha_i\}_i$. Another example is $\rho(U) := \left|\nabla U\right|$, which leads to TV regularization (section on p. 1008).

Reconstruction Operator

Assume that ρ is a sparsifying map for $U_{\text{true}} \in \mathscr{X}$ in (53) (or (54)). It is now natural to consider the most sparse ML solution (55) by solving (77) using a regularization functional given as the 0-norm of $U \mapsto \rho(U)$.

Much of the theoretical development in compressed sensing deals with when one can replace this 0-norm with a 1-norm, the latter being computationally feasible. Most of the mathematical results apply for linear finite dimensional reconstruction problems (54) where the noise in data is additive Gaussian. If the matrix representing the fully discretized forward operator (sensing matrix) fulfills certain mathematical criteria (restricted isometry property and coherence w.r.t. the sparsifying map), then the optimization below yields the sparsest solution to (54):

$$\min_{\boldsymbol{U} \in \mathbb{R}^n} \ \mu \left\| \rho(\boldsymbol{U}) \right\|_1 + \left\| \boldsymbol{T}(\boldsymbol{U}) - \boldsymbol{g} \right\|_{\mathbb{R}^m}^2 \quad \text{for } \boldsymbol{g} \in \mathbb{R}^m \text{ given as in (54).}$$

Adaptation to ET

Ideas of applying compressed sensing to ET have been discussed in [17] for STEM data from material science specimens. The forward operator is the ray transform and the data discrepancy functional is based on additive Gaussian noise model. The paper discusses dictionaries, like blobs (p. 1005), but examples show TV regularization.

Comments and Discussion

A key step in using sparsity is to suggest a sparsifying map. There is a wide range of (possibly over-complete) dictionaries, both analytic and learned, to choose from for sparsely representing texture/gray scale and edge information in images [48, 163]. Which ones are suitable for a specific ET application? As an example, the sparsest representation of a molecule is simply given by listing its atoms and their positions, e.g., following the RCSB Protein Data Bank specification. The corresponding sparsifying map is however only applicable when the electrostatic potential (the real part of the signal) is resolved to a 3D resolution of about 0.1 nm, which is way beyond what is reachable in ET applications in biology. Still, there are several notions for describing 3D structural information in macromolecular assemblies that could be used in the design of dictionaries. As an example, tertiary and quaternary structure descriptions are resolvable in ET, but it is unclear what dictionaries to use for encoding such structural information.

The next key issue is to what extent we have a sensing matrix that fulfills the criteria for replacing the 0-norm with the 1-norm. Unfortunately, there are no efficient methods for determining whether a sensing matrix is coherent w.r.t. a dictionary and/or fulfills the restricted isometry property. As an example, it is an open problem to provide deterministic and explicit methods that yield matrices that fulfill the restricted isometry property. On the other hand, it is well known that a random matrix will satisfy the restricted isometry property with high probability if its entries are samples of any sub-gaussian distribution [39, section 1.4.4]. There are also results in this direction for certain deterministic sensing matrices [21], but none of these results are applicable to ET.

Other Reconstruction Schemes

Here we very briefly mention other reconstruction methods used in ET. In [2] one applies geometric tomography to ET on STEM data from material sciences specimens. A similar approach is discrete tomography that was applied in [29] to ET on TEM data from biological specimens. This did not bring any benefits, but the situation is different in material sciences [12], see also comments regarding discrete ART on p. 1005.

Another approach deals with simultaneous reconstruction and segmentation within the framework of variational methods. Here, the Mumford–Shah functional can be used, see chapter ▸ Mumford and Shah Model and Its Applications to Image Segmentation and Image Restoration for an extensive review. In [94] one applies this approach to simulated ET data, and results are encouraging.

Finally, there is the possibility of making use of a priori shape information in the regularization (shape-based regularization). Shapes constitute the most important a priori knowledge that biologists use in their analysis of 3D reconstructions. A difficulty is that shapes are complex objects that are difficult to precisely describe as mathematical entities in a way feasible for computational treatment. Next, in biology no two objects have identical shape. Hence, the notion of shape must incorporate some form of statistics. The first approach to demonstrate the importance of shape information in the context of ET was taken in [69]. Here, variational regularization is used with a regularization functional that is based on a spatial prior that encodes local shape related geometric information. The interesting outcome of this attempt is that prescribing shape information on a small sub-set of the region of interest sharpens the entire reconstruction. An entirely different approach, that is much more robust against misspecification in a priori shape information, is based on shape metrics from the Large Deformation Diffeomorphic Metric Mapping (LDDM) framework in [194], see also chapter ▸ Shape Spaces. The idea is that one has shape information given as a set of regions $\mathfrak{O} := \{\Omega_j\}_{j=1,\ldots,k} \subset \Omega_0$ with associated *shape templates* $\mathfrak{I} := \{I_j\}_{j=1,\ldots,k}$ where $I_j : \Omega_j \to \mathbb{R}$ is such that

$$U_{\text{true}}\big|_{\Omega_j} \approx \phi_j . I_j \quad \text{for some admissible diffeomorphism } \phi_j \in G.$$

Here, G is the group of admissible diffeomorphisms that acts by the natural group action on the set $\mathscr{X}$ of potentials. The above simply encodes the assumption that shapes of certain substructures in the specimen, namely those contain in Ω_j, are known up to a deformation. Such information can be obtained from an initial 3D reconstruction, or by prior knowledge. Shape-based regularization is now defined as variational regularization (77) with $\mathcal{S}$ given as

$$\mathcal{S}(U) := \mathcal{S}_{\text{reg}}(U) + \alpha \mathcal{S}_{\text{shape}}(U; \mathfrak{O}, \mathfrak{I}, \boldsymbol{\mu}) \tag{86}$$

where $\mathcal{S}_{\text{reg}}$ encodes a priori regularity properties of U_{true} and $\mathcal{S}_{\text{shape}}$ is the *shape functional* accounting for shape information given by $\mathfrak{O}$ and $\mathfrak{I}$:

$$\mathcal{S}_{\text{shape}}\big(U; \mathfrak{O}, \mathfrak{I}, \boldsymbol{\mu}\big) := \min_{\phi_1,\ldots,\phi_k \in G} \sum_{j=1}^{k} \left[d_{\text{shape}}\big(U\big|_{\Omega_j}, I_j\big)^2 + \mu_j \left\| \phi_j . I_j - U\big|_{\Omega_j} \right\|_2 \right].$$

In the above, d_{shape} is the shape metric given by LDDM [194] that measures the shape similarity between the shape template and U_{true}. Hence, one combines an energy functional for regularity (like TV or Dirichlet energy) with a shape metric against a shape template [138]. Initial tests on 2D tomography (see Fig. 7) show that it is clear that a priori shape information, even imperfect, has a great potential in improving the reconstruction.

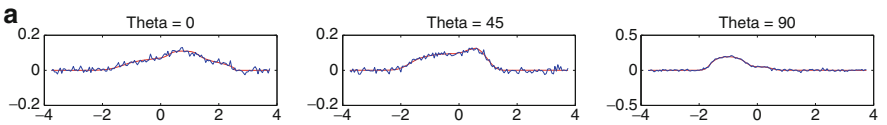
10 Validation

A serious obstacle in the development of computational methods for ET is the lack of proper validation tools. None of the reconstruction methods mentioned in this review has been validated mathematically (e.g., proof of convergence and error estimates) in a context that is relevant for ET. Moreover, there are no theoretical results that give bounds on the best possible "resolution," partly because it is unclear what notion of resolution to use. Hence, validation must be based on ET data from phantoms (specimens whose structure/potential is precisely known). Physical phantoms, i.e., physical specimens with a precisely known structure, only exist in material sciences. Hence, for many ET application one is confined to simulated data. To avoid committing an "inverse crime," both the simulator and phantom generator need to be as accurate as possible. This is nontrivial as explained in [166, 187].

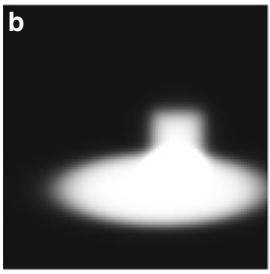
Finally, there is the issue of defining objective criteria for evaluating the quality of a reconstruction. This is closely related to the problem of defining a notion of resolution. Statistical approaches based on defining figures-of-merit are outlined in [30], but this approach only works for validation against simulate data.

Notion of Resolution

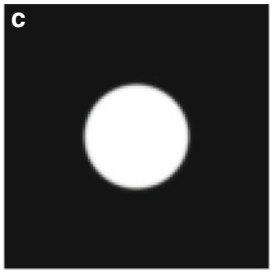
Intuitively one would expect that resolution *quantifies the size of the smallest features that can be reliably reconstructed*. Due to formal non-uniqueness (section "Incomplete Data, Uniqueness and Stability") in ET, one needs a notion of resolution for the visible wavefront set. This in turn requires precise control of the



Slightly noisy 2D parallel beam tomographic data of (b) from three directions, 0°, 45°, and 90°.

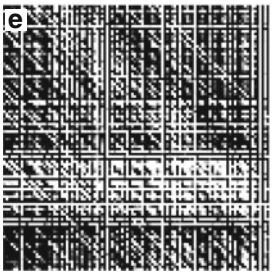


The phantom.

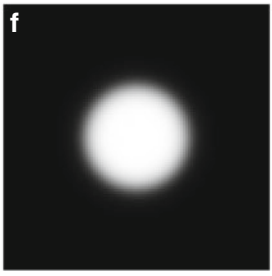


Shape template.

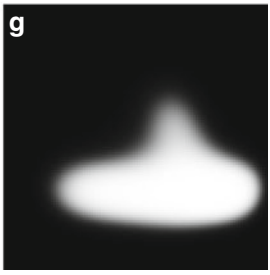
Shaped based regularization using (c).



FBP reconstruction.



Shape template (with smooth edges).



Shaped based regularization using (f).

Fig. 7 Test of 2D tomography with imperfect shape information. All reconstructions are based on data in (**a**). It is interesting to see how edge smoothness of the shape template affects the reconstruction. (**a**) Slightly noisy 2D parallel beam tomographic data of (**b**) from three directions, 0°, 45°, and 90°. (**b**) The phantom. (**c**) Shape template. (**d**) Shape-based regularization using (**c**). (**e**) FBP reconstruction. (**f**) Shape template (with smooth edges). (**g**) Shape-based regularization using (**f**)

constants in the Sobolev estimates that characterize the visible wavefront set. Some work in this direction is pursued in [153]. Secondly, it would also be of interest to have the corresponding characterization of visible wavefronts when the PSFs for the TEM optics and detector are included, and/or when on considers the full inverse scattering problem (i.e., the scattering operator is given by (6)).

The notions of resolution used in ET are all essentially different ways of estimating the spectral signal-to-noise ratio [144]. A serious drawback is that these notions do not account for the main degrading factor, the influence of the shot noise (section "Characteristics of the Noise"). Thus, a notion of resolution in ET has to

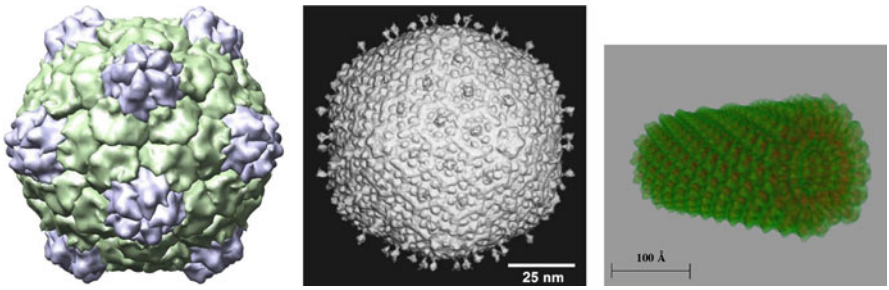


Fig. 8 Surface rendered electrostatic potentials calculated from atomic resolution models of a CPMV (*left*), a bacteriophage T4 head (*middle*), and a TMV (*right*) particle. The CPMV virion has a diameter of 28 nm [112], T4 bacteriophage head has a "diameter" of about 80 nm [108], and a TMV virion is about 300 nm long and 18 nm in diameter [132]. Hence, the diameter of the CPMV is about 1/3 of the diameter of the bacteriophage T4 head and 1.5 times the diameter of the TMV

include probabilistic concepts, see, e.g., [157, 180, 190] for ideas along these lines. To summarize, there is still no useful notion of resolution for ET.

11 Examples

This section shows results of different reconstructions methods applied to simulated and experimental single-axis tilt-series data, the latter courtesy of FEI. The WBP and SIRT reconstructions are obtained using IMOD v. 4.3.7 [124], the approximate inverse reconstructions are from Holger Kohr based on [102], and the variational regularization reconstructions are by software from Hans Rullgård based on [164].

Regularization Parameters

For WBP, it is the distance r (in pixels) of the radial reconstruction kernel in Fourier space before it is cut-off by a Gaussian with variance σ. For SIRT it is the number of iterations. For approximate inverse, it is the width γ of the Gaussian mollifier (paragraph on p. 999), and for variational methods, it is μ in (77).

Tilt-series data is not processed. WBP and SIRT assume data are samples of the ray transform, whereas the variational and approximate inverse methods implement (57) in which the PSF models the optics and the detector in the standard phase contrast model (section on p. 967).

Balls

This is a simulated single-axis tilt-series generated using the TEM simulation software in [166] of a phantom consisting of 40 balls with different size and contrast embedded in aqueous buffer.

Simulations represent a single-axis tilt-series acquired from a 300 keV conventional bright-field TEM. The tilt-angle ranges from $-60°$ to $60°$ with one

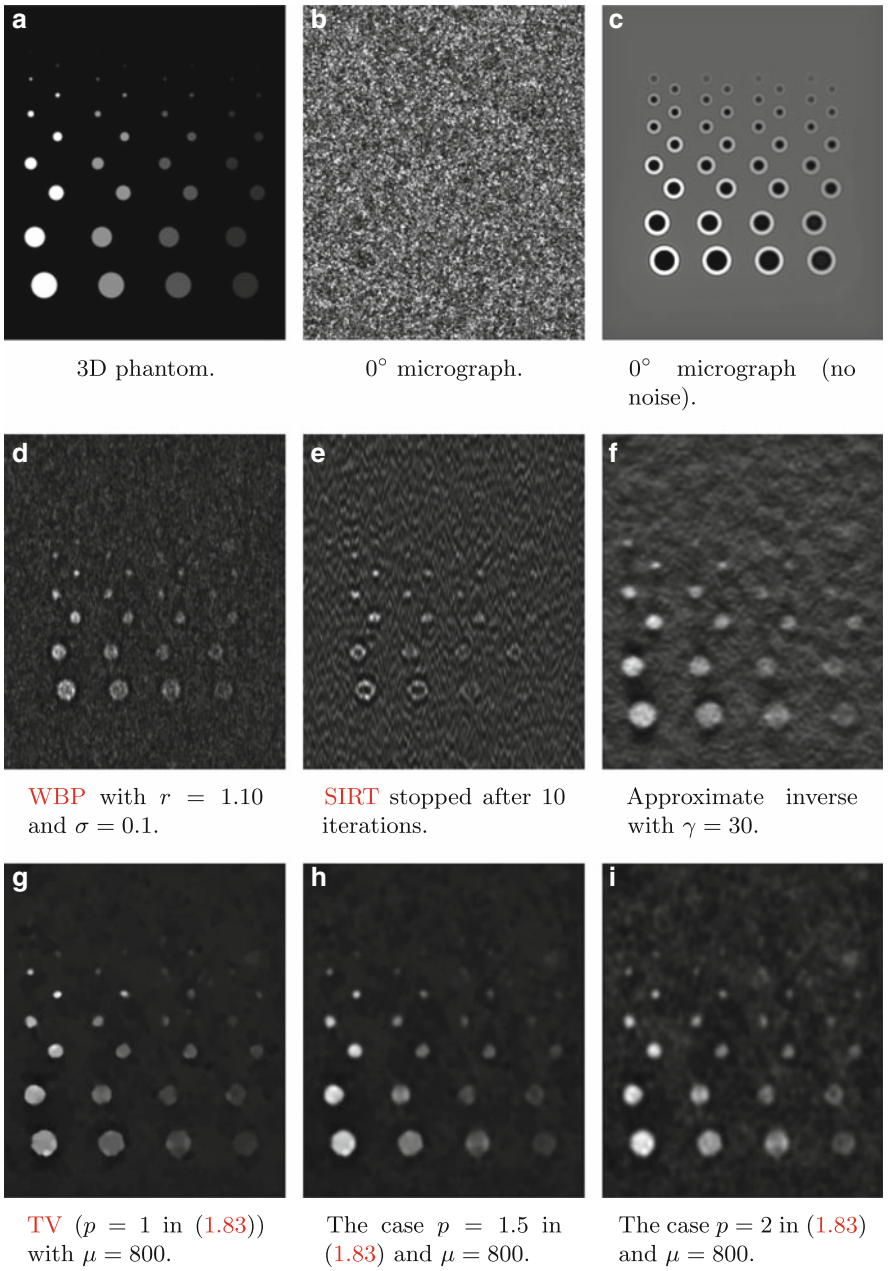
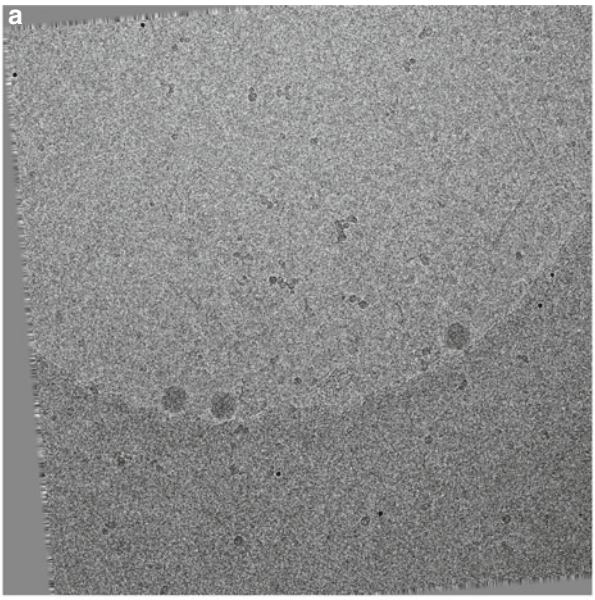
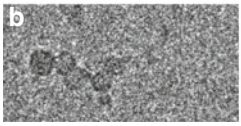


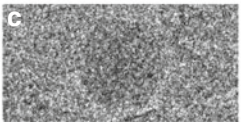
Fig. 9 Simulated data from a balls phantom. (**a**) shows a 2D cross-section of the phantom and (**b**) is the zero-tilt image, with (**c**) showing the noise-free version. The remaining images show the same 2D cross section as in (**a**) through different reconstructions. (**a**) 3D phantom. (**b**) 0° micrograph. (**c**) 0° micrograph (no noise). (**d**) WBP with $r = 1.10$ and $\sigma = 0.1$. (**e**) SIRT stopped after ten iterations. (**f**) Approximate inverse with $\gamma = 30$. (**g**) TV ($p = 1$ in (83)) with $\mu = 800$. (**h**) The case $p = 1.5$ in (83) and $\mu = 800$. (**i**) The case $p = 2$ in (83) and $\mu = 800$



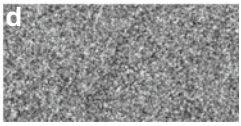
The 0.12° tilt micrograph, a 4096 × 4096 image in the aligned tilt-series.



Region 1 with CPMV.



Region 2 with bacteriophage T4 head.



Region 3 with TMV.

Fig. 10 Summary of experimental tilt-series. *Top image* (**a**) shows the zero-tilt micrograph and *bottom row* (**b**)–(**d**) show the three extracted regions that contain the particles illustrated in Fig. 8. (**a**) The 0.12° tilt micrograph, a 4096 × 4096 image in the aligned tilt-series. (**b**) Region 1 with CPMV. (**c**) Region 2 with bacteriophage T4 head. (**d**) Region 3 with TMV (Data courtesy of http://www.fei.com/)

Fig. 11 (continued) 2D slices through 3D reconstructions of the three regions of interest using different reconstruction methods. Images (**a**)–(**f**) are from region 1, (**g**)–(**l**) are from region 2, and (**m**)–(**r**) are from region 3. (**a**) WBP with $r = 1.1$ and $\sigma = 0.2$. (**b**) SIRT stopped after ten iterations. (**c**) Approximate inverse with $\gamma = 42$. (**d**) TV ($p = 1$ in (83)) with $\mu = 1{,}500$. (**e**) The case $p = 1.5$ in (83) and $\mu = 1{,}500$. (**f**) The case $p = 2$ in (83) and $\mu = 1{,}500$. (**g**) WBP with $r = 1.2$ and $\sigma = 0.4$. (**h**) SIRT stopped after ten iterations. (**i**) Approximate inverse with $\gamma = 42$. (**j**) TV ($p = 1$ in (83)) with $\mu = 1{,}500$. (**k**) The case $p = 1.5$ in (83) and $\mu = 1{,}500$. (**l**) The case $p = 2.0$ in (83) and $\mu = 1{,}500$. (**m**) WBP with $r = 1.01$ and $\sigma = 0.5$. (**n**) SIRT stopped after ten iterations. (**o**) Approximate inverse with $\gamma = 30$. (**p**) TV ($p = 1$ in (83)) with $\mu = 1{,}500$. (**q**) The case $p = 1.5$ in (83) and $\mu = 1{,}500$. (**r**) The case $p = 2.0$ in (83) and $\mu = 1{,}500$ (Data courtesy of http://www.fei.com/)

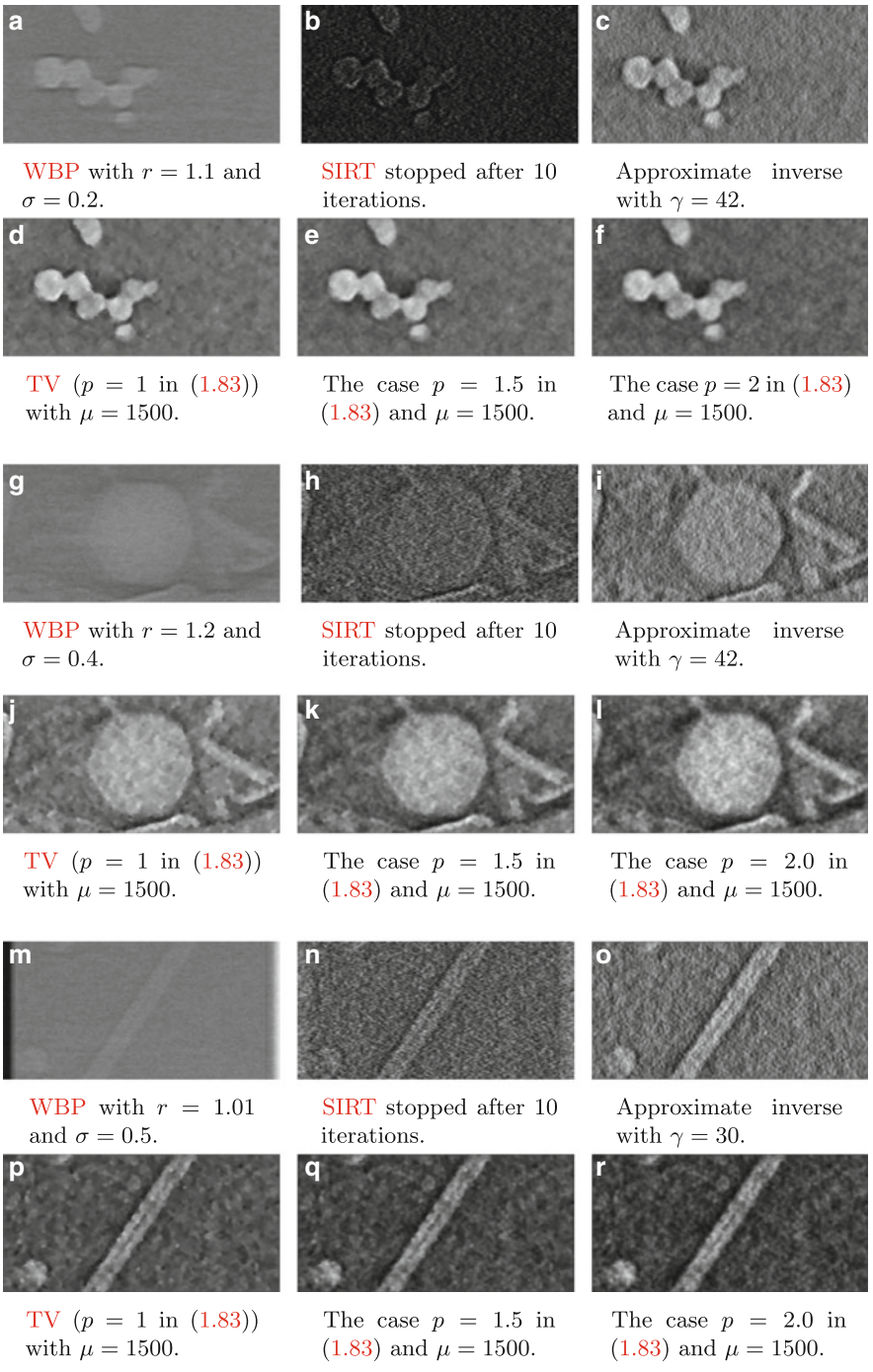

a
b
c
WBP with $r = 1.1$ and $\sigma = 0.2$.
SIRT stopped after 10 iterations.
Approximate inverse with $\gamma = 42$.
d
e
f
TV ($p = 1$ in (1.83)) with $\mu = 1500$.
The case $p = 1.5$ in (1.83) and $\mu = 1500$.
The case $p = 2$ in (1.83) and $\mu = 1500$.
g
h
i
WBP with $r = 1.2$ and $\sigma = 0.4$.
SIRT stopped after 10 iterations.
Approximate inverse with $\gamma = 42$.
j
k
l
TV ($p = 1$ in (1.83)) with $\mu = 1500$.
The case $p = 1.5$ in (1.83) and $\mu = 1500$.
The case $p = 2.0$ in (1.83) and $\mu = 1500$.
m
n
o
WBP with $r = 1.01$ and $\sigma = 0.5$.
SIRT stopped after 10 iterations.
Approximate inverse with $\gamma = 30$.
p
q
r
TV ($p = 1$ in (1.83)) with $\mu = 1500$.
The case $p = 1.5$ in (1.83) and $\mu = 1500$.
The case $p = 2.0$ in (1.83) and $\mu = 1500$.

Fig. 11 (continued)

micrograph every second degree, i.e., 61 micrographs in total. The 3D region of interest Ω_0 is a rectangular $210 \times 250 \times 40$ voxel region, with 0.5 nm voxels. Magnification is 25,000×, defocus is 3 μm, $f = 2.7$ mm (focal length), $C_s = 2.1$ mm (spherical aberration), $C_c = 2.2$ mm (chromatic aberration), aperture diameter is 40 μm, and $\alpha_c = 0.1$ mrad (condenser aperture angle). The detector has 16 μm pixels, $C_{\text{gain}} = 80$, and detector response follows (56) with $a = 0.7, b = 0.2, c = 0.1, \alpha = 10$, and $\beta = 40$. Finally, the total dose is 6,000 e^-/nm^2, which corresponds to 40 e^-/pixel in each micrograph.

Virions and Bacteriophages in Aqueous Buffer

This is an experimental single-axis tilt series of a cryo-fixated in vitro specimen that contains a mixture of TMV, Keyhole Limpet Hemocyanin, CPMV, and T4 bacteriophage particles in aqueous buffer, see Fig. 8.

The single-axis tilt-series was acquired by FEI using a 300 keV conventional bright-field TEM (FEI Titan Krios with a Falcon direct electron detector). It consists of 81 micrographs that were aligned using IMOD software [124]. There are three regions of interest listed below (voxel size is 0.4767 nm):

Region 1: $512 \times 256 \times 200$ containing mainly CPMV virions, see Fig. 10b for corresponding cut-out from the zero-tilt micrograph.
Region 2: $512 \times 256 \times 350$ containing mainly a single bacteriophage T4 head, see Fig. 10c for corresponding cut-out from the zero-tilt micrograph.
Region 3: $512 \times 256 \times 350$ containing mainly portion of a single TMV virion, see Fig. 10d for corresponding cut-out from the zero-tilt micrograph.

Micrographs are acquired at 29,370× magnification using a defocus of 6.0 μm and a detector with a pixel size of 14 μm. The total dose is unknown. So are the parameters for the TEM electron optical elements and illumination (Figs. 9, 10, and 11).

12 Conclusion

Much of the development in ET has circled around new improved and sample preparation. There has been tremendous progress in increasing the resolving power in electron optical elements, which in turn is important for imaging in material sciences. Current state-of-the-art TEMs have an optical resolving power of 0.05 nm [93], which is to be compared to the "size" of an isolated neutral atom that ranges between 0.03 and 0.3 nm. There are also strong indications that we are reaching the physical limits for the optical resolving power of an electron microscope, and that is not due to imperfections in optics but rather "noise" from the specimen [158]. Thus, it is reasonable to assume that the optical resolving power will not improve much beyond 0.05 nm.

For ET in life sciences, the aforementioned development is not that relevant since the resolving power of the TEM optics does not limit the 3D "resolution." Instead, the limitation here is due to the noise in image data, which in turn is due to the limited dose (section "The Dose Problem"). On the other hand, technological development regarding automation in sample preparation and data collection will have a big impact on ET in life sciences. TEMs are now stable enough to allow for automated recording of multiple tilt-series from multiple regions of interests within one or multiple specimens. Thus, the amount of available high-quality tilt-series is rapidly increasing and mathematics will have a key role in providing better 3D reconstructions as well as extracting useful information from such 3D reconstructions.

The inverse problem in ET also contains open mathematical problems, e.g., there are open problems related to uniqueness and stability (section "Incomplete Data, Uniqueness, and Stability"). Furthermore, the design of regularization methods need to better account for the regularity/sparsity that a specimen poses. An example is the variants of TV type of regularization which need to be analyzed carefully from this viewpoint. Notions of sparsity and shape that are applicable to flexible molecular assemblies and/or subcellular structures is a central theme. Yet another central topic is the regularization parameter selection in problems with highly noisy data and/or complex noise models (Poisson and Gaussian, or correlated Poisson and Gaussian). Another area where much remains to be done is to mathematically analyze intertwined reconstruction schemes. These are approaches for handling nuisance parameters (paragraph on p. 989) in the context of classical regularization that often work surprisingly well. Finally, the high level of noise in ET indicates that the best framework for reconstruction is offered by statistical regularization. This framework is yet to be applied to ET in the sense that one recovers not only a single estimator but also a measure of the uncertainty. Statistical regularization is computationally demanding, so it has had limited applications to imaging problems. Nevertheless, it can be successfully applied to real 3D tomography problems [182], so the approach should also be applicable to ET.

Acknowledgments The writing of this review chapter would not be possible without the help of several people. Jan Boman at the Department of Mathematics, Stockholm University and Todd Quinto at the Department of Mathematics, Tufts University provided valuable advice and help, especially regarding material in section "Analytic Methods." Hans Rullgård at Comsol has provided advice and support regarding the usage of the TEM simulation and the TV-regularization softwares used in Sect. 11. Remco Schoenmakers at FEI generously provided the data used in section "Virions and Bacteriophages in Aqueous Buffer" as well as access to the image in Fig. 1. Milos Vulevic and Bernd Reiger in TU Delft provided valuable insight into the detector modeling in subsection on p. 963, Sergej Masich at the Department of Cell and Molecular Biology, Karolinska Institutet helped out in the alignment of the tilt-series in section "Virions and Bacteriophages in Aqueous Buffer" as well as in running the IMOD reconstructions in Sect. 11. Holger Kohr and Alfred Louis at the Department of Mathematics, Saarlands University contributed to the material on the approximate inverse method and phase contrast tomography. Kohr also provided the approximate inverse reconstructions in Sect. 11. Finally, Günther Uhlmann at the Department of Mathematics, Washington State University provided valuable insight into uniqueness and stability issues discussed in section "Electron–Specimen Interaction."

Work on this chapter is financially supported by the Swedish Foundation for Strategic Research.

Cross-References

- ▸ Compressive Sensing
- ▸ EM Algorithms
- ▸ EM Algorithms from a Non-stochastic Perspective
- ▸ Inverse Scattering
- ▸ Iterative Solution Methods
- ▸ Microlocal Analysis in Tomography
- ▸ Mumford and Shah Model and Its Applications to Image Segmentation and Image Restoration
- ▸ Numerical Methods and Applications in Total Variation Image Restoration
- ▸ Shape Spaces
- ▸ Tomography
- ▸ Total Variation in Imaging
- ▸ Wave Phenomena

References

1. Aganj, I., Bartesaghi, A., Borgnia, M., Liao, H.Y., Sapiro, G., Subramaniam, S.: Regularization for Inverting the Radon Transform with Wedge Consideration. IMA Preprint Series, vol. 2144. Institute for Mathematics and its Applications, Minneapolis (2006)
2. Alpers, A., Gardner, R.J., König, S., Pennington, R.S., Boothroyd, C.B., Houben, L., Dunin-Borkowski, R.E., Batenburg, K.J.: Geometric reconstruction methods for electron tomography. Ultramicroscopy **128**, 42–54 (2013)
3. Amat, F., Castanõ-Diez, D., Lawrence, A., Moussavi, F., Winkler, H., Horowitz, M.: Alignment of cryo-electron tomography datasets. In: Jensen, G.J. (ed.) Cryo-EM, Part B: 3-D Reconstruction. Methods in Enzymology, Chap. 13, vol. 482. pp. 343–367. Academic, San Diego (2010)
4. Ammari, H., Bahouri, H., Dos Santos Ferreira, D., Gallagher, I.: Stability estimates for an inverse scattering problem at high frequencies. J. Math. Anal. Appl. **400**, 525–540 (2013)
5. Andersen, A.H., Kak, A.C.: Simultaneous algebraic reconstruction technique (SART): a superior implementation of the ART algorithm. Ultrason. Imaging **6**(1), 81–94 (1984)
6. Asoubar, D., Zhang, S., Wyrowski, F., Kuhn, M.: Parabasal field decomposition and its application to non-paraxial propagation. Opt. Exp. **20**(21), 23502–23517 (2012)
7. Ayache, J., Beaunier, L., Boumendil, J., Ehret, G., Laub, D.: Sample Preparation Handbook for Transmission Electron Microscopy: Methodology. Springer, New York (2010)
8. Ayache, J., Beaunier, L., Boumendil, J., Ehret, G., Laub, D.: Sample Preparation Handbook for Transmission Electron Microscopy: Techniques. Springer, New York (2010)
9. Baker, L.A., Rubinstein, J.L.: Radiation damage in electron cryomicroscopy. In: Jensen, G.J. (ed.) Cryo-EM, Part A: Sample Preparation and Data Collection. Methods in Enzymology, Chap. 15, vol. 481, pp. 371–388. Academic, San Diego (2010)
10. Bakushinsky, A.: Remarks on choosing a regularization parameter using the quasi-optimality and ratio criterion. USSR Comput. Math. Math. Phys. **24**(4), 181–182 (1984)
11. Bakushinsky, A.B., Yu, M.: Kokurin. Iterative Methods for Approximate Solution of Inverse Problem. Mathematics and its Applications, vol. 577. Springer, Dordrecht (2004)
12. Batenburg, K.J., Kaiser, U., Kübel, C.: 3D imaging of nanomaterials by discrete tomography. Microsc. Microanal. **12**, 1568–1569 (2006). Extended abstract of a paper presented at Microscopy and Microanalysis 2006 in Chicago, IL, USA, 30 July–3 August, 2006

13. Batenburg, K.J., Bals, S., Sijbers, J., Kübel, C., Midgley, P.A., Hernandez, J.C., Kaiser, U., Encina, E.R., Coronado, E.A., Van Tendeloo, G.: 3D imaging of nanomaterials by discrete tomography. Ultramicroscopy **109**, 730–740 (2009)
14. Benning, M., Brune, C., Burger, M., Müller, J.: Higher-order tv methods—enhancement via Bregman iteration. J. Sci. Comput. **54**(2–3), 269–310 (2013)
15. Benvenuto, F., La Camera, A., Theys, C., Ferrari, A., Lantéri, H., Bertero, M.: The study of an iterative method for the reconstruction of images corrupted by Poisson and Gaussian noise. Inverse Prob. **24**, 035016 (20 pp.) (2008)
16. Betero, M., Lantéri, H., Zanni, L.: Iterative image reconstruction: a point of view. In: Censor, Y., Jiang, M., Louis, A.K. (eds.) Mathematical Methods in Biomedical Imaging and Intensity-Modulated Radiation Therapy (IMRT). Publications of the Scuola Normale Superiore: CRM Series, vol. 7, pp. 37–63. Springer, Pisa (2008)
17. Binev, P., Dahmen, W., DeVore, R., Lamby, P., Savu, D., Sharpley, R.: Compressed sensing and electron microscopy. In: Vogt, T., Dahmen, W., Binev, P. (eds.) Modeling Nanoscale Imaging in Electron Microscopy, Nanostructure Science and Technology, pp. 73–126. Springer, New York (2012)
18. Blomgren, P., Chan, T., Mulet, P., Wong, C.K.: Total variation image restoration: numerical methods and extensions. In Proceedings of the 1997 IEEE International Conference on Image Processing, Santa Barbara, vol. 3, pp. 384–387 (1997)
19. Böhm, J., Frangakis, A.S., Hegerl, R., Nickell, S., Typke, D., Baumeister, W.: Toward detecting and identifying macromolecules in a cellular context: template matching applied to electron tomograms. Proc. Natl. Acad. Sci. **97**(26), 14245–14250 (2000)
20. Boman, J., Quinto, E.T.: Support theorems for real analytic Radon transforms on line complexes in r^3. Trans. Am. Math. Soc. **335**, 877–890 (1993)
21. Bourgain, J., Dilworth, S., Ford, K., Konyagin, S., Kutzarova, D.: Explicit constructions of RIP matrices and related problems. Duke Math. J. **159**(1), 145–185 (2011)
22. Burvall, A., Lundström, U., Takman, P.A.C., Larsson, D.H., Hertz, H.M.: Phase retrieval in x-ray phase-contrast imaging suitable for tomography. Opt. Exp. **19**(11), 10359–10376 (2011)
23. Busch, H.: Berechnung der Bahn von Kathodenstrahlen im axialsymmetrischen elektromagnetischen Felde. Ann. Phys. **386**, 974–993 (1926)
24. Busch, H.: Über die Wirkungsweise der Konzentrierungsspule bei der Braunschen Rohre. Electr. Eng. (Archiv fur Elektrotechnik) **18**, 583–594 (1927)
25. Byrne, C.L.: Applied Iterative Methods. A. K. Peters, Wellesley (2008)
26. Cai, Y., Zhao, Y., Tang, Y.: Exponential convergence of a randomized Kaczmarz algorithm with relaxation. In: Gaol, F.L., Nguyen, Q.V. (eds.) Proceedings of the 2011–2nd International Congress on Computer Applications and Computational Science. Volume 2. Advances in Intelligent and Soft Computing, vol. 145, pp. 467–473 (2012)
27. Candès, E.J., Romberg, J., Tao, T.: Stable signal recovery from incomplete and inaccurate measurements. Commun. Pure Appl. Math. **59**(8), 1207–1223 (2006)
28. Carazo, J.-M., Carrascosa, J.L.: Restoration of direct Fourier three-dimensional reconstructions of crystalline specimens by the method of convex projections. J. Microsc. **145**(2), 159–177 (1987)
29. Carazo, J.-M., Sorzano, C.O.S., Reitzel, E., Schröder, R., Marabini, R.: Discrete tomography in electron microscopy. In: Herman, G.T., Kuba, A. (eds.) Discrete Tomography. Foundations, Algorithms and Applications, Chap. 18, pp. 405–416. Birkhäuser, Boston (1999)
30. Carazo, J.-M., Herman, G.T., Sorzano, C.O.S., Marabini, R.: Algorithms for three-dimensional reconstruction from the imperfect projection data provided by electron microscopy. In: Frank, J. (ed.) Electron Tomography: Methods for Three-Dimensional Visualization of Structures in the Cell, Chap. 7, 2nd edn., pp. 217–243. Springer, Boston (2006)
31. Chadan, K.: Inverse problems in potential scattering. In: Chadan, K., Colton, D., Päivärinta, L., Rundell, W. (eds.) An Introduction to Inverse Scattering and Inverse Spectral Problems. SIAM Monographs on Mathematical Modeling and Computation, Chap. 4, vol. 2. SIAM, Philadelphia (1997)

32. Chakrabarti, A., Zickler, T.: Depth and deblurring from a spectrally-varying depth-of-field. In: Proceedings of the European Conference on Computer Vision 2012. Springer, New York (2012)
33. Chan, T.F., Esedoglu, S., Park, F., Yip, A.: Recent developments in total variation image restoration. In: Paragios, N., Chen, Y., Faugeras, O. (eds.) Handbook of Mathematical Models in Computer Vision. Springer, New York (2005)
34. Çınlar, E.: Probability and Stochastics. Graduate Texts in Mathematics, vol. 261. Springer, New York (2011)
35. Colton, D., Kress, R.: Inverse scattering. In: Scherzer, O. (ed.) Handbook of Mathematical Methods in Imaging, Chap. 13, pp. 551–598. Springer, Berlin (2011)
36. Colton, D., Kress, R.: Inverse Acoustic and Electromagnetic Scattering Theory. Applied Mathematical Sciences, vol. 93, 3rd edn. Springer, New York (2013)
37. Cruz-Uribe, D.V., Fiorenza, A.: Variable Lebesgue Spaces: Foundations and Harmonic Analysis. Applied and Numerical Harmonic Analysis. Springer, Basel (2013)
38. Csiszar, I.: Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems. Ann. Stat. **19**(4), 2032–2066 (1991)
39. Davenport, M., Duarte, M., Eldar, Y., Kutyniok, G.: Introduction to compressed sensing. In: Eldar, Y.C., Kutyniok, G. (eds.) Compressed Sensing: Theory and Applications, Chap. 1, pp. 1–65. Cambridge University Press, Cambridge (2011)
40. Davidson, M.E.: The ill-conditioned nature of the limited angle tomography problem. SIAM J. Appl. Math. **43**(2), 428–448 (1983)
41. De Rosier, D.J., Klug, A.: Reconstruction of three dimensional structures from electron micrographs. Nature **217**, 130–134 (1968)
42. Defrise, M., Clack, R., Townsend, D.W.: Image reconstruction from truncated, two-dimensional, parallel projections. Inverse Probl. **11**(2), 287–313 (1995)
43. DeRosier, D.J.: The reconstruction of three-dimensional images from electron micrographs. Contemp. Phys. **12**(5), 437–452 (1971)
44. Ebanks, B., Sahoo, P., Sander, W.: Characterization of Information Measures. World Scientific, Singapore (1998)
45. Egerton, R.F.: Electron Energy-Loss Spectroscopy in the Electron Microscope, 3rd edn. Springer, New York (2011)
46. Egerton, R.F., Li, P., Malac, M.: Radiation damage in the TEM and SEM. Micron **35**, 399–409 (2004)
47. Eggermont, P.P.B.: Maximum entropy regularization for Fredholm integral equations of the first kind. SIAM J. Math. Anal. **24**(6), 1557–1576 (1993)
48. Elad, M.: Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing. Springer (2010)
49. Engl, H.W., Landl, G.: Convergence rates for maximum entropy regularization. SIAM J. Numer. Anal. **30**(5), 1509–1536 (1993)
50. Engl, H.W., Hanke, M., Neubauer, A.: Regularization of Inverse Problems. Mathematics and its Applications, vol. 375. Kluwer Academic Publishers (2000)
51. Fanelli, D., Öktem, O.: Electron tomography: a short overview with an emphasis on the absorption potential model for the forward problem. Inverse Probl. **24**(1), 013001 (51 pp.) (2008)
52. Faridani, A.: Introduction to the mathematics of computed tomography. In: Uhlmann, G. (ed.) Inside Out: Inverse Problems and Applications. MSRI Publications, vol. 47, pp. 1–46. Cambridge University Press, Cambridge (2003)
53. Faridani, A.: Fan-beam tomography and sampling theory. In: Ólafsson, G., Quinto, E.T. (eds.) The Radon Transform, Inverse Problems, and Tomography. Proceedings of Symposia in Applied Mathematics, vol. 63, pp. 43–66. American Mathematical Society, Providence (2006)
54. Faridani, A., Finch, D., Ritman, E.L., Smith, K.T.: Local tomography II. SIAM J. Appl. Math. **57**(4), 1095–1127 (1997)

55. Faruqi, A.R., McMullan, G.: Electronic detectors for electron microscopy. Q. Rev. Biophys. **44**(3), 357–390 (2011)
56. Felea, R., Quinto, E.T.: The microlocal properties of the local 3-D SPECT operator. SIAM J. Math. Anal. **43**(3), 1145–1157 (2011)
57. Fernández, J.J., Li, S., Crowther, R.A.: CTF determination and correction in electron cryotomography. Ultramicroscopy **106**(7), 587–596 (2006)
58. Foley, J.T., Butts, R.R.: Uniqueness of phase retrieval from intensity measurements. J. Opt. Soc. Am. A **71**(8), 1008–1014 (1981)
59. Foucart, S., Rahut, H.: A Mathematical Introduction to Compressive Sensing. Applied and Numerical Harmonic Analysis. Springer, New York (2013)
60. Frank, J.: Three-Dimensional Electron Microscopy of Macromolecular Assemblies, 2nd edn. Oxford University Press, Oxford (2006)
61. Frank, J.: Single-particle reconstruction of biological macromoleculses in electron micrscopy – 30 years. Q. Rev. Biophys. **42**, 139–158 (2009)
62. Frikel, J., Quinto, T.: Characterization and reduction of artifacts in limited angle tomography. Inverse Probl. **29**(12), 125007 (2013)
63. Fultz, B., Howe, J.: Transmission Electron Microscopy and Diffractometry of Materials. Graduate Texts in Physics, 4th edn. Springer, Berlin (2013)
64. Gardueno, E., Herman, G.T.: Optimization of basis functions for both reconstruction and visualization. Discrete Appl. Math. **139**, 95–111 (2004)
65. Gilbert, P.: Iterative methods for the three-dimensional reconstruction of an object from projections. J. Theor. Biol. **36**(1), 105–117 (1972)
66. Gilbert, R., Hackl, K., Xu, Y.: Inverse problem for wave propagation in a perturbed layered half-space. Math. Comput. Model. **45**(1–2), 21–33 (2007)
67. Gil-Rodrigo, E., Portilla, J., Miraut, D., Suarez-Mesa, R.: Efficient joint poisson-gauss restoration using multi-frame l2-relaxed-l0 analysis-based sparsity. In: 18th IEEE International Conference on Image Processing (ICIP), 2011, pp. 1385–1388 (2011)
68. Glaeser, R.M., Downing, K.H., DeRozier, D., Chu, W., Frank, J.: Electron Crystallography of Biological Macromolecules. Oxford University Press, Oxford (2006)
69. Gopinath, A., Xu, G., Ress, D., Öktem, O., Subramaniam, S., Bajaj, C.: Shape-based regularization of electron tomographic reconstruction. IEEE Trans. Med. Imaging **31**(12), 2241–2252 (2012)
70. Gordon, R., Bender, R., Herman, G.T.: Algebraic reconstruction techniques (ART) for three-dimensional electron microscopy and X-ray photography. J. Theor. Biol. **29**(3), 471–481 (1970)
71. Goris, B., Van den Broek, W., Batenburg, K.J., Mezerji, H.H., Bals, S.: Electron tomography based on a total variation minimization reconstruction technique. Ultramicroscopy **113**, 120–130 (2012)
72. Goris, B., Roelandts, T., Batenburg, K.J., Mezerji, H.H., Bals, S.: Advanced reconstruction algorithms for electron tomography: from comparison to combination. Ultramicroscopy **127**, 40–47 (2013)
73. Greenleaf, A., Uhlmann, G.: Non-local inversion formulas for the X-ray transform. Duke Math. J. **58**, 205–240 (1989)
74. Greenleaf, A., Uhlmann, G.: Composition of some singular Fourier integral operators and estimates for restricted X-ray transforms. Annales de l'Institut Fourier **40**, 443–466 (1990)
75. Greenleaf, A., Uhlmann, G.: Microlocal techniques in integral geometry. In: Grinberg, E., Quinto, E.T. (eds.) Integral Geometry and Tomography. Proceedings of the AMS-IMS-SIAM Joint Summer Research Conference. Contemporary Mathematics, vol. 113, pp. 121–136. American Mathematical Society, Providence (1990)
76. Guillemin, V., Sternberg, S.: Geometric Asymptotics. Mathematical Surveys and Monographs, vol. 14. American Mathematical Society, Providence (1977). Revised edition (June 1990) edition

77. Hansen, P.-C.: Rank-Deficient and Discrete Ill-Posed Problems: Numerical Aspects of Linear Inversion. SIAM Monographs on Mathematical Modeling and Computation, vol. 4. SIAM, Philadelphia (1997)
78. Hawkes, P.W.: The electron microscope as a structure projector. In: Frank, J. (ed.) Electron Tomography. Methods for Three-Dimensional Visualization of Structures in the Cell, Chap. 3, 2nd edn., pp. 83–111. Springer, New York (2006)
79. Hawkes, P.W., Kasper, E.: Principles of Electron Optics. Wave Optics, vol. 3. Academic, San Diego (1995)
80. Hawkes, P.W., Kasper, E.: Principles of Electron Optics. Applied Geometrical Optics, vol. 2. Academic, San Diego, (1996)
81. Herman, G.T.: Fundamentals of Computerized Tomography: Image Reconstruction from Projections. Advances in Computer Vision and Pattern Recognition, 2nd edn. Springer, New York (2010)
82. Hermann, U., Noll, D.: Adaptive image reconstruction using information measures. SIAM J. Control Optim. **38**(4), 1223–1240 (2000)
83. Hohage, T., Werner, F.: Iteratively regularized Newton-type methods for general data misfit functionals and applications to Poisson data. Numerische Mathematik, **123**(4), 745–779 (2013)
84. Hohage, T., Werner, F.: Convergence rates for inverse problems with impulsive noise. SIAM J. Numer. Anal. **52**(3), 1203–1221 (2014)
85. Hoppe, W., Langer, R., Knesch, G., Poppe, C.: Protein-kristallstrukturanalyse mit elektronenstrahlen. Naturwissenschaften **55**, 333–336 (1968)
86. Hörmander L.: Fourier integral operators I. Acta Math. **127**(1–2), 79–183 (1971)
87. Isakov, V.: Increased stability in the continuation for the Helmholtz equation with variable coefficient. In: Ancona, F., Lasiecka, I., Littman, W., Triggiani, R. (eds.) Control Methods in PDE-Dynamical Systems. Contemporary Mathematics, vol. 426, pp. 243–254. American Mathematical Society, Providence (2007)
88. Ivanyshyn, O., Kress, R.: Inverse scattering for surface impedance from phase-less far field data. J. Comput. Phys. **230**, 3443–3452 (2011)
89. Jin, S., Yang, X.: Computation of the semiclassical limit of the Schrödinger equation with phase shift by a level set method. J. Sci. Comput. **35**(2), 144–169 (2008)
90. Jonas, P., Louis, A.K.: Phase contrast tomography using holographic measurements. Inverse Probl. **20**(1), 75–102 (2004)
91. Kaipio, J., Somersalo, E.: Statistical and Computational Inverse Problems. Applied Mathematical Sciences, vol. 160. Springer, New York (2005)
92. Kaltenbacher, B., Neubauer, A., Scherzer, O.: Iterative Regularization Methods for Nonlinear Ill-Posed Problems. Radon Series on Computational and Applied Mathematics, vol. 6. Walter de Gruyter, Berlin (2008)
93. Kisielowski, C., Freitag, B., Bischoff, M., van Lin, H., Lazar, S., Knippels, G., Tiemeijer, P., van der Stam, M., von Harrach, S., Stekelenburg, M., Haider, M., Müller, H., Hartel, P., Kabius, B., Miller, D., Petrov, I., Olson, E., Donchev, T., Kenik, E.A., Lupini, A., Bentley, J., Pennycook, S., Minor, A.M., Schmid, A.K., Duden, T., Radmilovic, V., Ramasse, Q., Erni, R., Watanabe, M., Stach, E., Denes, P., Dahmen, U.: Detection of single atoms and buried defects in three dimensions by aberration-corrected electron microscopy with 0.5 Å information limit. Microsc. Microanal. **14**, 454–462 (2008)
94. Klann, E.: A Mumford–Shah-like method for limited data tomography with an application to electron tomography. SIAM J. Imaging Sci. **4**(4), 1029–1048 (2011)
95. Klibanov, M.V.: Phaseless inverse scattering problems in 3-d. Technical Report March (2013). arXiv:1303.0923v1, arXiv
96. Klibanov, M.V.: Uniqueness of two phaseless inverse acoustics problems in 3-d. Technical Report March (2013). arXiv:1303.7384v1, arXiv
97. Klug, A., DeRosier, D.J.: Reconstruction of three dimensional structures from electron micrographs. Nature **217**, 130–134 (1968)

98. Knoll, M., Ruska, E.: Das Elektronenmikroskop. Z. Phys. A: Hadrons Nucl. **78**, 318–339 (1932)
99. Koch, C.T., Lubk, A.: Off-axis and inline electron holography: a quantitative comparison. Ultramicroscopy **110**(5), 460–471 (2010)
100. Kohl, H., Rose, H.: Theory of image formation by inelastically scattered electrons in the electron microscope. Adv. Electr. Electron Phys. **65**, 173–227 (1985)
101. Kohr, H.: Fast and high-quality reconstructions in electron tomography. In: Simos, T.E., Psihoyios, G., Tsitouras, Ch. (eds.) ICNAAM 2010: International Conference of Numerical Analysis and Applied Mathematics 2010, 19–25 September 2010, Rhodes (Greece). AIP Conference Proceedings, vol. 1281, pp. 1979–1981 (2010)
102. Kohr, H., Louis, A.K.: Fast and high-quality reconstruction in electron tomography based on an enhanced linear forward model. Inverse Probl. **27** 045008 (20 pp.) (2011)
103. Koster, A.J., Bárcena, M.: Cryotomography: Low-dose automated tomography of frozen-hydrated specimens. In: Frank, J. (ed.) Electron Tomography. Methods for Three-Dimensional Visualization of Structures in the Cell, Chap. 4, 2nd edn., pp. 113–161. Springer, New York (2006)
104. Krupchyk, K., Lassas, M., Uhlmann, G.: Inverse problems with partial data for a magnetic Schrödinger operator in an infinite slab and on a bounded domain. Commun. Math. Phys. **312**, 87–126 (2012)
105. Lassas, M., Saksman, E., Siltanen, S.: Discretization-invariant bayesian inversion and besov space priors. Inverse Probl. Imaging **3**(1), 87–122 (2009)
106. Lawrence, A., Bouwer, J.C., Perkins, G.A., Ellisman, M.H.: Transform-based backprojection for volume reconstruction of large format electron microscope tilt series. J. Struct. Biol. **154**, 144–167 (2006)
107. Lee, H., Rose, Z., Hambach, H., Wachsmuth, P., Kaiser, U.: The influence of inelastic scattering on EFTEM images – exemplified at 20 kV for graphene and silicon. Ultramicroscopy **134**, 102–112 (2013)
108. Leimana, P.G., Kanamaru, S., Mesyanzhinov, V.V., Arisaka, F., Rossmann, M.G.: Structure and morphogenesis of bacteriophage T4. Cell. Mol Life Sci. **60**, 2356–2370 (2003)
109. Lewitt, R.M.: Alternatives to voxels for image representation in iterative reconstruction algorithms. Phys. Med. Biol. **37**(3), 705–716 (1992)
110. Li, J., Shen, Z., Yin, R., Zhang, X.: A reweighted ℓ^2-method for image restoration with poisson and mixed Poisson–Gaussian noise. UCLA Computational and Applied Mathematics Reports 12–84. Department of Mathematics University of California, Los Angeles (2012)
111. Lin, P.D.: New Computation Methods for Geometrical Optics. Springer Series in Optical Sciences, vol. 178. Springer (2014)
112. Lin, T., Chen, Z., Usha, R., Stauffacher, C.V., Dai, J.-B., Schmidt, T., Johnson, J.E.: The refined crystal structure of cowpea mosaic virus at 2.8 A resolution. Virology **265**, 20–34 (1999)
113. Liu, H., Ralston, J., Runborg, O., Tanushev, N.M.: Gaussian beam methods for the helmholtz equation (2013). arXiv [math.NA] 1304.1291v1, arXiv
114. Louis, A.K.: Incomplete data problems in x-ray computerized tomography. Numer. Math. **48**(3), 251–262 (1986)
115. Louis, A.K.: A unified approach to regularization methods for linear ill-posed problems. Inverse Probl. **15**(2), 489–498 (1999)
116. Louis, A.K.: Development of algorithms in computerized tomography. In: Ólafsson, G., Quinto, E.T. (eds.) The Radon Transform, Inverse Problems, and Tomography. Proceedings of Symposia in Applied Mathematics, vol. 63, pp. 25–42. American Mathematical Society, Providence (2006)
117. Louis, A.K.: Feature reconstruction in inverse problems. Inverse Probl. **27**(6), 065010 (2011)
118. Louis, A.K., Maass, P.: A mollifier method for linear operator equations of the first kind. Inverse Probl. **6**(3), 427–440 (1990)
119. Macaulay, V.A., Buck, B.: Linear inversion by the method of maximum entropy. Inverse Probl. **5**, 859–874 (1989)

120. Marabini, R., Herman, G.T., Carazo, J.-M.: 3D reconstruction in electron microscopy using ART with smooth spherically symmetric volume elements (blobs). Ultramicroscopy **72**, 53–65 (1998)
121. Marburg, S.: Discretization requirements: how many elements per wavelength are necessary? In: Marburg, S., Nolte, B. (eds.) Computational Acoustics of Noise Propagation in Fluids – Finite and Boundary Element Methods, Chap. 11, pp. 309–332. Springer, Berlin (2008)
122. Markoe, A.: Analytic Tomography. Encyclopedia of Mathematics and its Applications, vol. 106. Cambridge University Press, Cambridge (2006)
123. Marone, F., Münch, B., Stampanoni, M.: Fast reconstruction algorithm dealing with tomography artifacs. In: Stock, S.R. (ed.) Developments in X-Ray Tomography VII, San Diego, CA, 1 August 2010. Proceedings of SPIE, vol. 7804, pp. 780410-1–780410-11 (2010)
124. Mastronarde, D.N.: Dual-axis tomography: an approach with alignment methods that preserve resolution. J. Struct. Biol. **120**, 343–352 (1997)
125. Matsushima, K., Schimmel, H., Buehling, S., Wyrowski, F.: Propagation of electromagnetic fields between nonparallel planes. In: Wyrowski, F. (ed.) Wave-Optical Systems Engineering II. Proceedings of SPIE, vol. 5182, pp. 55–62. SPIE, Bellingham (2003)
126. Midgley, P.A., Weyland, M.: 3D electron microscopy in the physical sciences. The development of Z-contrast and EFTEM tomography. Ultramicroscopy **96**, 413–431 (2003)
127. Midgley, P.A., Weyland, M., Stegmann, H.: Applications of electron tomography. In: Banhart, J. (ed.) Advanced Tomographic Methods in Materials Research and Engineering. Monographs on the Physics and Chemistry of Materials, Chap. 12, vol. 66, pp. 335–372. Oxford University Press, Oxford (2008)
128. Muga, J.G., Palao, J.P., Navarro, B., Egusquiza, I.L.: Complex absorbing potentials. Phys. Rep. **395**, 357–426 (2004)
129. Müller, H.: A coherence function approach to image simulation. Ph.D. thesis, Fachbereich Physik, Technischen Universität Darmstadt, Darmstadt, Germany (2000)
130. Myers, G.R., Gureyev, T.E., Paganin, D.M.: Stability of phase-contrast tomography. J. Opt. Soc. Am. A: Opt. Image Sci. Vis. **24**(9), 2516–2526 (2007)
131. Nagayasu, S., Uhlmann, G., Wang, J.-N.: Increasing stability in an inverse problem for the acoustic equation. Inverse Probl. **29**(2), 025012 (11 pp.) (2013)
132. Namba, K., Pattanayek, R., Stubbs, G.: Visualization of protein-nucleic acid interactions in a virus. Refined structure of intact tobacco mosaic virus at 2.9 Å resolution by X-ray fiber diffraction. J. Mol. Biol. **208**(2), 307–325 (1989)
133. Narasimha, R., Aganj, I., Bennett, A., Borgnia, M.J., Zabransky, D., Sapiro, G., McLaughlin, S.W., Milne, J.L.S., Subramaniam, S.: Evaluation of denoising algorithms for biological electron tomography. J. Struct. Biol. **164**(1), 7–17 (2008)
134. Natterer, F.: The Mathematics of Computerized Tomography. Classics in Applied Mathematics, vol. 32. SIAM, Philadelphia (2001)
135. Natterer, F.: An error bound for the Born approximation. Inverse Probl. **20**, 447–452 (2004)
136. Natterer, F., Wübbeling, F.: Mathematical Methods in Image Reconstruction. SIAM Monographs on Mathematical Modeling and Computation, vol. 5. SIAM, Philadelphia (2001)
137. Noll, D.: Consistency of a nonlinear deconvolution method with applications in image restoration. Adv. Math. Sci. Appl. **7**(2), 789–808 (1997)
138. Öktem, O., Bajaj, C., Ravikumar, P.: Summary of results regarding shape based regularization for de-noising and 2D tomography. Preliminary Progress Report (2013)
139. Paganin, D., Mayo, S.C., Gureyev, T.E., Miller, P.R., Wilkins, S.W.: Simultaneous phase and amplitude extraction from a single defocused image of a homogeneous object. J. Microsc. **206**(Part 1), 33–40 (2002)
140. Palamodov, V.P.: Stability in diffraction tomography and a nonlinear "basic theorem". J. d'Analyse Mathématique **91**(1), 247–268 (2003)
141. Palamodov, V.P.: Reconstructive Integral Geometry. Monographs in Mathematics, vol. 98. Birkhäuser, Basel (2004)
142. Palamodov, V.P.: Inverse scattering as nonlinear tomography. Wave Motion **47**, 635–640 (2010)

143. Penczek, P.A.: Fundamentals of three-dimensional reconstruction from projections. In: Jensen, G.J. (ed.) Cryo-EM, Part B: 3-D Reconstruction. Methods in Enzymology, Chap. 1, vol. 482, pp. 1–33. Academic, San Diego (2010)
144. Penczek, P.A.: Resolution measures in molecular electron microscopy. In: Jensen, G.J. (ed.) Cryo-EM, Part B: 3-D Reconstruction. Methods in Enzymology, Chap. 3, vol. 482, pp. 73–100. Academic, San Diego (2010)
145. Penczek, P.A., Frank, J.: Resolution in electron tomography. In: Frank, J. (ed.) Electron Tomography. Methods for Three-Dimensional Visualization of Structures in the Cell, Chap. 10, 2nd edn., pp. 307–330. Springer, New York (2006)
146. Peng, L.-M., Dudarev, S.L., Whelan, M.J.: High-Energy Electron Diffraction and Microscopy. Monographs on the Physics and Chemistry of Materials, vol. 61, Oxford University Press, Oxford (2004)
147. Pereyra, V.: Ray tracing methods for inverse problems. Inverse Probl. **16**(6), R1–R35 (2000)
148. Plitzko, J.M., Baumeister, W.: Cryoelectron tomography (CET). In: Hawkes, P.W., Spence, J.C.H. (eds.) Science of Microscopy, Chap. 7, pp. 535–604. Springer (2008)
149. Pohjola, V.: An Inverse Boundary Value Problem for the Magnetic Schrödinger Operator on a Half Space. Licentiate thesis. Department of Mathematics and Statistics, University of Helsinki (2012)
150. Quinto, E.T.: Singularities of the X-ray transform and limited data tomography in r^2 and r^3. SIAM J. Math. Anal. **24**, 1215–1225 (1993)
151. Quinto, E.T., Öktem, O.: Local tomography in electron microscopy. SIAM J. Appl. Math. **68**(5), 1282–1303 (2008)
152. Quinto, E.T., Rullgård, H.: Electron microscope tomography over curves. Oberwolfach Reports 18/2010, Mathematisches Forschungsinstitut Oberwolfach (2010)
153. Quinto, E.T., Rullgård, H.: Local sobolev estimates of a function by means of its Radon transform. Inverse Probl. Imaging **4**(4), 721–734 (2010)
154. Quinto, E.T., Rullgård, H.: Local singularity reconstruction from integrals over curves in r^3. Inverse Probl. Imaging **7**(2), 585–609 (2013)
155. Quinto, E.T., Skoglund, U., Öktem, O.: Electron lambda-tomography. Proc. Natl. Acad. Sci. **106**(51), 21842–21847 (2009)
156. Radermacher, M.: Weighted back-projection methods. In: Frank, J. (ed.) Electron Tomography. Methods for Three-Dimensional Visualization of Structures in the Cell, Chap. 8, 2nd edn., pp. 245–273. Springer, New York (2006)
157. Ram, S., Ward, S.E., Ober, R.J.: A stochastic analysis of distance estimation approaches in single molecule microscopy: quantifying the resolution limits of photon-limited imaging systems. Multidimension. Syst. Signal Process. **24**, 503–542 (2013)
158. Reich, E.S.: Imaging hits noise barrier: physical limits mean that electron microscopy may be nearing highest possible resolution. Nature **499**(7457), 135–136 (2013)
159. Resmerita, E.: Regularization of ill-posed problems in Banach spaces: convergence rates. Inverse Probl. **21**, 1303–1314 (2005)
160. Rose, H.: Information transfer in transmission electron microscopy. Ultramicroscopy **15**(3), 173–192 (1984)
161. Rose, H.: Advances in electron optics. In: Ernst, F., Rühle, M. (eds.) High-Resolution Imaging and Spectrometry of Materials. Springer Series in Materials Science, Chap. 5, pp. 189–270. Springer, Berlin (2003)
162. Rose, H.: Geometrical Charged-Particle Optics. Springer Series in Optical Sciences, vol. 142, 2nd edn. Springer, New York (2012)
163. Rubinstein, R., Bruckstein, A.M., Elad, M.: Dictionaries for sparse representation modeling. Proc. IEEE **98**(6), 1045–1057 (2010)
164. Rullgård, H.: A new principle for choosing regularization parameter in certain inverse problems. Department of Mathematics, Stockholm University (2008). arXiv report in math.NA arXiv:0803.3713v2

165. Rullgård, H., Öktem, O., Skoglund, U.: A componentwise iterated relative entropy regularization method with updated prior and regularization parameter. Inverse Prob. **23**, 2121–2139 (2007)
166. Rullgård, H., Öfverstedt, L.-G., Masich, S., Daneholt, B., Öktem, O.: Simulation of transmission electron microscope images of biological specimens. J. Microsc. **243**(3), 234–256 (2011)
167. Runborg, O.: Mathematical models and numerical methods for high frequency waves. Commun. Comput. Phys. **2**(5), 827–880 (2007)
168. Sato, M.: Hyperfunctions and partial differential equations. In: Proceedings of the 2nd Conference on Functional Analysis and Related Topics, Tokyo, pp. 91–94. Tokyo University, Tokyo University Press, Tokyo (1969)
169. Scherzer, O., Grasmair, M., Grossauer, H., Haltmeier, M., Lenzen, F.: Variational Methods in Imaging. Applied Mathematical Sciences, vol. 167. Springer, New York (2009)
170. Schuster, T.: The Method of Approximate Inverse: Theory and Applications. Lecture Notes in Mathematics, vol. 1906. Springer, Heidelberg (2007)
171. Schuster, T., Kaltenbacher, B., Hofmann, B., Kazimierski, K.S.: Regularization Methods in Banach Spaces. Radon Series on Computational and Applied Mathematics, vol. 10. Walter de Gruyter, Berlin (2012)
172. Skoglund, U., Öfverstedt, L.-G., Burnett, R.M., Bricogne, G.: Maximum-entropy three-dimensional reconstruction with deconvolution of the contrast transfer function: a test application with Adenovirus. J. Struct. Biol. **117**, 173–188 (1996)
173. Smith, K.T.: Inversion of the X-ray transform. In: McLaughlin, D.W. (ed.) Inverse Problems. SIAM-AMS Proceedings, vol. 14, pp. 41–52. American Mathematical Society, Providence (1984)
174. Snyder, D.L., Schutz, T.J., O'Sullivan, J.A.: Deblurring subject to nonnegative constraint. IEEE Trans. Signal Process. **40**, 1143–1150 (1992)
175. Sorzano, C.O.S., Otero, A., Olmos, E.M., Carazo, J.-M.: Error analysis in the determination of the electron microscopical contrast transfer function parameters from experimental power spectra. BMC Struct. Biol. **9**(18) (2009)
176. Spence, J.C.H.: High-Resolution Electron Microscopy. Monographs on the Physics and Chemistry of Materials, vol. 60, 3rd edn. Oxford University Press, New York (2003)
177. Turonŏvá, B.: Simultaneous Algebraic Reconstruction Technique for Electron Tomography Using OpenCL. MSc thesis. Saarland University, Faculty of Natural Sciences and Technology I, Department of Computer Science (2011)
178. Vainberg, E.I., Kazak, I.A., Kurozaev, V.P.: Reconstruction of the internal three-dimensional structure of objects based on real-time integral projections. Sov. J. Nondestruct. Test. **17**, 415–423 (1981)
179. Vainshtein, B.K., Barynin, V.V., Gurskaya, G.V.: The hexagonal crystalline structure of catalase and its molecular structure. Dokl. Akad. Nauk SSSR **182**, 569–572 (1968). See also Sov. Phys. Dokl. **13**, 838–841 (1969)
180. Van Aert, S., den Dekker, A.J., Van Dyck, D., van den Bos, A.: The notion of resolution. In: Hawkes, P.W., Spence, J.C.H. (eds.) Science of Microscopy, Chap. 20, pp. 1228–1265. Springer, New York (2007)
181. Van den Broek, W., Koch, C.T.: General framework for quantitative three-dimensional reconstruction from arbitrary detection geometries in TEM. Phys. Rev. B: Condens. Matter Mater. Phys. **87**(18), 184108 (11 pp.) (2013)
182. Vanska, S., Lassas, M., Siltanen, S.: Statistical X-ray tomography using empirical Besov priors. Int. J. Tomography Stat. **30**, 3–32 (2009)
183. Verbeeck, J., Schattschneider, P., Rosenauer, A.: Image simulation of high resolution energy filtered TEM images. Ultramicroscopy **109**, 350–360 (2009)
184. Voortman, L.M., Stallinga, S., Schoenmakers, R.H., van Vliet, L.J., Rieger, B.: A fast algorithm for computing and correcting the CTF for tilted, thick specimens in TEM. Ultramicroscopy **111**(8), 1029–1036 (2011)

185. Voortman, L.M., Franken, E.M., van Vliet, L.J., Rieger, B.: Fast, spatially varying CTF correction in TEM. Ultramicroscopy **118**, 26–34 (2012)
186. Vulović, M., Franken, E.M., Ravelli, R.B., van Vliet, L.J., Rieger, B.: Precise and unbiased estimation of astigmatism and defocus in transmission electron microscopy. Ultramicroscopy **116**, 115–134 (2012)
187. Vulović, M., Ravelli, R.B., van Vliet, L.J., Koster, A.J., Lazić, I., Lücken, U., Rullgård, H., Öktem, O., Rieger, B.B.: Image formation modeling in cryo-electron microscopy. J. Struct. Biol. **183**(1), 19–32 (2013)
188. Vulović, M., Rieger, B., van Vliet, L.J., Koster, A.J., Ravelli, R.B.: A toolkit for the characterization of CCD cameras for transmission electron microscopy. Acta Crystallogr. Sect. D: Biol. Crystallogr. **66**, 97–109 (2010)
189. Vulović, M., Voortman, L.M., van Vliet, L.J., Rieger, B.: When to use the projection assumption and the weak-phase object approximation in phase contrast cryo-EM. Ultramicroscopy **136**, 61–66 (2014)
190. Wang, A., Turner, S., Van Aert, S., Van Dyck, D.: An alternative approach to determine attainable resolution directly from HREM images. Ultramicroscopy **133**, 50–61 (2013)
191. Xiong, Q., Morphew, M.K., Schwartz, C.L., Hoenger, A.H., Mastronarde, D.N.: CTF determination and correction for low dose tomographic tilt series. J. Struct. Biol. **168**, 378–387 (2009)
192. Xu, W., Xu, F., Jones, M., Keszthelyi, B., Sedat, J., Agard, D., Mueller, K.: High-performance iterative electron tomography reconstruction with long-object compensation using graphics processing units (gpus). J. Struct. Biol. **171**(2), 142–153 (2010)
193. Xu, G., Li, M., Gopinath, A., Bajaj, C.: Inversion of electron tomography images using l^2-gradient flows – computational methods. J. Comput. Math. **29**(5), 501–525 (2011)
194. Younes, L.: Shapes and Diffeomorphisms. Applied Mathematical Sciences, vol. 171. Springer, New York (2010)
195. Yserentant, H.: Regularity and Approximability of Electronic Wave Functions. Lecture Notes in Mathematics, vol. 2000. Springer, Berlin (2010)
196. Zeidler, E.: Nonlinear Functional Analysis and Its Applications. Part 3: Variational Methods and Optimization. Springer, New York (1985)
197. Zuo, J.M.: Electron detection characteristics of a slow-scan CCD camera, imaging plates and film, and electron image restoration. Microsc. Res. Tech. **49**, 245–268 (2000)

Optical Imaging

Simon R. Arridge, Jari P. Kaipio, Ville Kolehmainen, and Tanja Tarvainen

Contents

1 Introduction 1034
2 Background 1035
Spectroscopic Measurements 1035
Imaging Systems 1036
3 Mathematical Modeling and Analysis 1037
Radiative Transfer Equation 1037
Diffusion Approximation 1039
Hybrid Approaches Utilizing the DA 1044
Green's Functions and the Robin to Neumann Map 1045
The Forward Problem 1046
Schrödinger Form 1047
Perturbation Analysis 1048
Linearization 1052
Adjoint Field Method 1054
Light Propagation and Its Probabilistic Interpretation 1056
4 Numerical Methods and Case Examples 1059
Image Reconstruction in Optical Tomography 1059
Bayesian Framework for Inverse Optical Tomography Problem 1060
Experimental Results 1067
5 Conclusion 1075
Cross-References 1075
References 1075

S.R. Arridge (✉)
Department of Computer Science, University College London, London, UK
e-mail: S.Arridge@cs.ucl.ac.uk

J.P. Kaipio
Department of Mathematics, University of Auckland, Auckland, New Zealand
e-mail: jari@math.auckland.ac.nz

V. Kolehmainen • T. Tarvainen
Department of Physics and Mathematics, University of Eastern Finland, Kuopio, Finland
e-mail: Ville.Kolehmainen@uef.fi; Tanja.Tarvainen@uef.fi

© Springer Science+Business Media New York 2015

O. Scherzer (ed.), *Handbook of Mathematical Methods in Imaging*,
DOI 10.1007/978-1-4939-0790-8_21

Abstract

This chapter discusses diffuse optical tomography. We present the origins of this method in terms of spectroscopic analysis of tissue using near-infrared light and its extension to an imaging modality. Models for light propagation at the macroscopic and mesoscopic scale are developed from the radiative transfer equation (RTE). Both time- and frequency-domain systems are discussed. Some formal results based on Green's function models are presented, and numerical methods are described based on discrete finite element method (FEM) models and a Bayesian framework for image reconstruction. Finally, some open questions are discussed.

1 Introduction

Optical imaging in general covers a wide range of topics. In this chapter, we mean techniques for *indirect* imaging using light as a method for obtaining observations of a subject. In a typical experiment, a highly scattering medium is illuminated by a narrow collimated beam, and the light that propagates through the medium is collected by an array of detectors. There are many variants of this basic scenario. For instance, the source may be pulsed or time harmonic, coherent, or incoherent, and the illumination may be spatially structured or multispectral. Likewise, the detector may be time or frequency resolved, polarization or phase sensitive, located in the near or far field, and so on. The inverse problem that is considered is to reconstruct the optical properties of the medium from boundary measurements. The mathematical formulation of the corresponding forward problem is dictated primarily by *spatial scale*, ranging from the Maxwell equations at the microscale to the radiative transport equation at the mesoscale and to the diffusion theory at the macroscale. In addition, experimental time scales vary from the femtosecond on which light pulses are generated, through the nanosecond on which diffuse waves propagate, to the millisecond scale on which biological activation takes place and still longer for pathophysiologic changes.

In this chapter, we concentrate primarily on the macroscopic scale and the diffusion model for light propagation. The derivation of this model and its limits of applicability are discussed in section "Radiative Transfer Equation." Historically, a large amount of early development considered analytic forms for the Green's function of the diffusion equation and series expressions for the effect of perturbations of these propagators by inhomogeneities; usually, only first-order linear methods were considered. These are discussed in section "Green's Functions and the Robin to Neumann Map." As computational methods become more readily available, more sophisticated approaches using optimization and Bayesian methods are becoming more accepted. We discuss these approaches in Sect. 4.

2 Background

Figure 1 schematically illustrates the two main types of measurement system: time resolved and intensity modulated. In the former, a short duration pulse ~5–10 ps is employed, and in the latter, a steady-state intensity is created, modulated at a frequency in the range 100–1,000 MHz. Obviously, the spectrum of frequencies in the time domain is many order higher than in the frequency-domain systems themselves, although the higher frequencies are very heavily damped and carry no information. A third domain is "DC" systems – these are the same as frequency domain, without the modulation. They are much simpler and cheaper, but without a complex wave, the inverse problem is nonunique [6].

Spectroscopic Measurements

Attenuation of light in the near infrared (NIR) is due to absorption and scattering. The parameter of most interest is absorption which is caused by chromophores of variable concentration such as hemoglobin in its oxygenated and deoxygenated states. In the absence of scattering, the change in light intensity obeys the Beer–Lambert law

$$-\ln \frac{I_{\text{in}}}{I_{\text{out}}} = \mu_{\text{a}}\, d = \alpha_c [c] d, \tag{1}$$

where d is the source-detector separation, which is equal to the optical path length, $[c]$ is the concentration of chromophore c, and α_c is the absorption coefficient per unit length per unit concentration of chromophore c and can usually be obtained in vitro. In the presence of scattering, the optical path length of transmitted photons

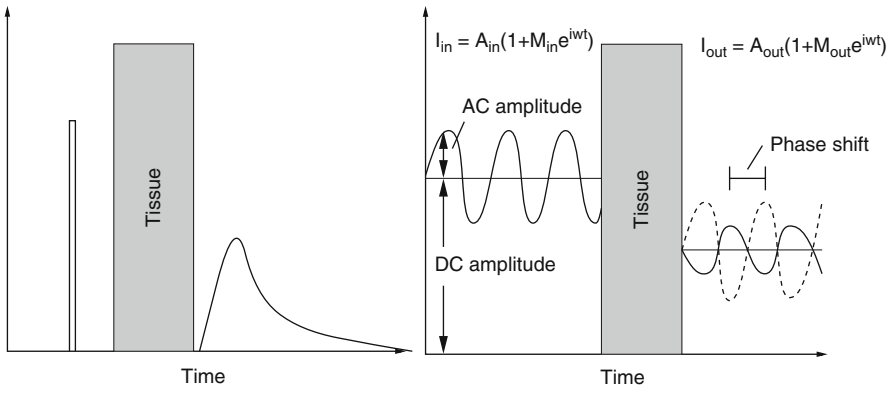


Fig. 1 Optical transillumination measurements made with a time-resolved system (*left*) or an intensity-modulated system (*right*)

follows a much more complex relationship. Hence, attenuation measurements alone do not allow quantification of chromophore concentration.

Continuous intensity (DC) instruments measure changes in the intensity of light leaving the tissue surface [56]. This is frequently done in a purely spectroscopic manner, i.e., to obtain only global changes in chromophore concentration. In order to quantify concentration changes, additional information is required. One approach is to derive an approximate *differential path length factor* (DPF), which restores the approximate Beer–Lambert law for small changes in concentration

$$-\delta \ln \frac{I_{\text{in}}}{I_{\text{out}}} = \text{DPF}\, \delta \mu_{\text{a}}\, d. \tag{2}$$

Since there are typically several contributing chromophores, light of different wavelengths in the NIR region is employed, and regression techniques are used to find their relative weightings [26]. It was shown empirically [29] that the DPF is simply the mean time of light multiplied by the speed of light in the tissue. In fact, this relationship follows naturally from the diffusion approximation of light transport [8]. Furthermore, it is equally well approximated by the change in phase of an intensity-modulated system, at least at low modulation frequencies.

Intensity-modulated measurements were first reported by [69]. Most systems use a heterodyne technique to mix the transmitted light with a reference beam of slightly different modulation frequency, thus producing a lower frequency envelope that is easier to detect using RF equipment. Time-resolved systems were first developed using a *streak camera* [29, 49, 73], an instrument with exceptionally high time resolution in the picosecond range but with high cost, relatively low dynamic range, and a significant inherent temporal nonlinearity due to a sinusoidal ramp voltage. Alternatively, *time-correlation single photon counting* (TCSPC) systems measure arrival times of individual photons by comparison with a reference pulse using a time-to-amplitude converter (TAC) device [20, 79, 83]. These systems have a high dynamic range and excellent temporal linearity.

Imaging Systems

Imaging methods can be divided into *direct* systems which seek to detect heterogeneities in tissue by analyzing the transmitted (or, in some cases, reflected) light and *indirect* systems which attempt to solve the inverse problem of image reconstruction. The latter is the main emphasis of this article although the former is historically the precedent, in a similar manner in which x-ray radiographs were the precursor to x-ray computed tomography (CT).

Transillumination of candle light for a patient suffering from hydrocephalus was reported as early as 1831, but the first significant attempt at diagnostic imaging using optical radiation was for breast lesions and was made by Cutler [27], who used a lamp held under the breast in a darkened room. However, even at this stage, multiple scattering effects caused a notable degradation in image quality. The recognition

of this fact led to many attempts to eliminate or minimize the degradation due to scattering ranging from collimation [55] and polarization discrimination [84] to coherence gating using holographic gating [90] or heterodyne detection [82].

With the introduction of time-resolved detectors came the natural attempt to use temporal gating to discriminate early arriving photons (which necessarily have the shortest optical path and therefore suffer the least number of scatterings) from later arriving photons which have undergone multiple scatterings and therefore have ill-determined photon paths. The early implementations of this idea used a Kerr gate as an ultrafast shutter [99]. However, this technique is limited to relatively low-scattering media due to the small dynamic range of the Kerr shutter. Other studies have been based on the streak camera [42] or TCSPC [19] systems described in section "Spectroscopic Measurements."

The attempt to physically discriminate between photons that have undergone different numbers of scattering events is inherently limited by the statistical likelihood of the low scattering number photons arriving at the detector. For the relatively optically thick tissues that are of interest in breast cancer screening or brain imaging, these photons are overwhelmed by noise. For this reason, indirect methods that solve an inverse problem based on recovering the spatially varying optical parameters that provide the best fit of a photon transport model with the measured data are becoming more prevalent. Within this framework, the three basic strategies (time resolved, intensity modulated, and DC systems) have all been developed and reported. In addition, many different geometrical arrangements have been investigated. Initial studies have been on 2D slice-by-slice imaging, although it is apparent that the photon propagation must in reality be described by a 3D model. Fully 3D methods are now appearing.

In the remainder of this article, we will discuss the inverse problem and the strategies that have been adopted in order to solve it. In order to analyze this problem, we first have to consider the model of photon transport in dense media.

3 Mathematical Modeling and Analysis

Radiative Transfer Equation

In optical imaging, light transport through a medium containing scattering particles is described by transport theory [54]. In transport theory, the particle conservation within a small volume element of phase space is investigated. The wave phenomenon of particles is ignored. The transport theory can be modeled through stochastic methods and deterministic methods. In the stochastic approach, individual particle interactions are modeled as the particles are scattered and absorbed within the medium. The two stochastic methods that have been used in optical imaging are the Monte Carlo method and the random walk theory, of which two, the Monte Carlo is the most often used [5].

In deterministic approach, particle transport is described with integrodifferential equations which can be solved either analytically or numerically [5]. In optical

imaging, a widely accepted model for light transport is the radiative transport equation (RTE). The RTE is a one-speed approximation of the transport equation, and thus it basically assumes that the energy (or speed) of the particles does not change in collisions and that the refractive index is constant within the medium. For discussion of photon transport in medium with spatially varying refractive index, see, e.g., [16, 36, 62, 72].

Let $\Omega \subset \mathbb{R}^n$, $n = 2$ or 3 denote the physical domain where n is the dimension of the domain. The medium is considered isotropic in the sense that the probability of scattering between two directions depends only on the relative angle between those directions and not on an absolute direction. For discussion of light propagation in anisotropic medium, see, e.g., [44]. Furthermore, let $\partial\Omega$ denote the boundary of the domain and $\hat{s} \in S^{n-1}$ denote a unit vector in the direction of interest. The RTE is written in time domain as

$$\begin{aligned} &\frac{1}{c}\frac{\partial \phi(r,\hat{s})}{\partial t} + \hat{s}\cdot\nabla\phi(r,\hat{s}) + (\mu_s + \mu_a)\phi(r,\hat{s}) \\ &\quad = \mu_s \int_{S^{n-1}} \Theta(\hat{s}\cdot\hat{s}')\phi(r,\hat{s}')\mathrm{d}\hat{s}' + q(r,\hat{s}) \end{aligned} \tag{3}$$

and in frequency domain as

$$\begin{aligned} &\frac{\mathrm{i}\omega}{c}\phi(r,\hat{s}) + \hat{s}\cdot\nabla\phi(r,\hat{s}) + (\mu_s + \mu_a)\phi(r,\hat{s}) \\ &\quad = \mu_s \int_{S^{n-1}} \Theta(\hat{s}\cdot\hat{s}')\phi(r,\hat{s}')\mathrm{d}\hat{s}' + q(r,\hat{s}), \end{aligned} \tag{4}$$

where c is the speed of light in the medium, i is the imaginary unit, ω is the angular modulation frequency of the input signal, and $\mu_s = \mu_s(r)$ and $\mu_a = \mu_a(r)$ are the scattering and absorption coefficients of the medium, respectively. The scattering coefficient represents the probability per unit length of a photon being scattered, and the absorption coefficient represents the probability per unit length of a photon being absorbed. Furthermore, $\phi(r,\hat{s})$ is the radiance, $\Theta(\hat{s}\cdot\hat{s}')$ is the scattering phase function, and $q(r,\hat{s})$ is the source inside Ω. The radiance can be defined such that the amount of power transfer in the infinitesimal angle $\mathrm{d}\hat{s}$ in direction $\hat{s}$ at time t through an infinitesimal area $\mathrm{d}S$ is given by

$$\phi(r,\hat{s};t)\hat{s}\cdot\hat{\nu}\mathrm{d}S\mathrm{d}\hat{s},$$

where $\hat{\nu}$ is the normal to the surface $\mathrm{d}S$ [54]. The scattering phase function $\Theta(\hat{s}\cdot\hat{s}')$ describes the probability that a photon with an initial direction $\hat{s}'$ will have a direction $\hat{s}$ after a scattering event. In optical imaging, the most usual phase function for isotropic material is the Henyey–Greenstein scattering function [47] which is of the form

$$\Theta(\hat{s}\cdot\hat{s}') = \begin{cases} \frac{1}{2\pi}\frac{1-g^2}{(1+g^2-2g\hat{s}\cdot\hat{s}')}, & n=2, \\ \frac{1}{4\pi}\frac{1-g^2}{(1+g^2-2g\hat{s}\cdot\hat{s}')^{3/2}}, & n=3, \end{cases} \tag{5}$$

where g is the scattering shape parameter that defines the shape of the probability density and it gets values between $-1 < g < 1$. With the value $g = 0$, the scattering probability density is a uniform distribution. For forward dominated scattering, $g > 0$, and for backward dominated scattering, $g < 0$. The time-domain and frequency-domain representations of the RTE are related through Fourier transform.

In order to obtain a unique solution for the RTE, the ingoing radiance distribution on the boundary $\partial\Omega$, that is, $\phi(r,\hat{s})$ for $\hat{s}\cdot\hat{\nu} < 0$, where $\hat{\nu}$ is the outward unit normal, needs to be known [24]. Several boundary conditions can be applied to the RTE [1, 33, 57]. In optical imaging, the boundary condition which assumes that no photons travel in an inward direction at the boundary $\partial\Omega$ is used [5]

$$\phi(r,\hat{s}) = 0, \quad r \in \partial\Omega, \quad \hat{s}\cdot\hat{n} < 0. \tag{6}$$

This boundary condition, also known as the free surface boundary condition and the vacuum boundary condition, implies that once a photon escapes the domain Ω, it does not reenter it. The boundary condition (6) can be modified to include a boundary source $\phi_0(r,\hat{s})$ at the source position $\varepsilon_j \subset \partial\Omega$, and it can be written in the form [96]

$$\phi(r,\hat{s}) = \begin{cases} \phi_0(r,\hat{s}), & r \in \cup_j \varepsilon_j, \quad \hat{s}\cdot\hat{n} < 0 \\ 0, & r \in \partial\Omega \setminus \cup_j \varepsilon_j, \quad \hat{s}\cdot\hat{n} < 0. \end{cases} \tag{7}$$

In optical imaging, the measurable quantity is the exitance $J_n(r)$ on the boundary of the domain. It is defined as [5]

$$J_n(r) = \int_{S^{n-1}} (\hat{s}\cdot\hat{\nu})\phi(r,\hat{s})\mathrm{d}\hat{s}, \quad r \in \partial\Omega. \tag{8}$$

Diffusion Approximation

In optical imaging, light propagation in tissues is usually modeled with the diffusion approximation (DA) to the RTE. The most typical approach to derive the DA from the RTE is to expand the radiance, the source term, and the phase function into series using the spherical harmonics and truncate the series [5, 24, 33]. If the spherical harmonics series is truncated at the Nth moment, P_N approximation is obtained [5, 33]. The first-order spherical harmonics approximation is referred to as the P_1 approximation, and the DA can be regarded as a special case for that. The most typical approach for utilizing the P_N approximations in optical imaging has been to use them in angular discretization of the numerical solution of the RTE [14, 100].

An alternative to the P_N approximation is the Boltzmann hierarchy approach, in which moments of radiance are used to form a set of coupled equations that approximate the RTE [38]. Furthermore, the DA can be derived using asymptotic techniques [7, 17] leading to generalized diffusion equation or by using projection algebra [33, 57]. If the speed of light is not constant, a diffusion equation with spatially varying indices of refraction can be derived [16].

Here, a short review of the derivation of the DA is given according to [54, 57]. First, the P_1 approximation is derived, and then, the DA is formed as a special case for that. In the DA framework, the radiance is approximated by

$$\phi(r,\hat{s}) \approx \frac{1}{|S^{n-1}|}\Phi(r) + \frac{n}{|S^{n-1}|}\hat{s}\cdot J(r), \tag{9}$$

where $\Phi(r)$ and $J(r)$ are the photon density and photon current which are defined as

$$\Phi(r) = \int_{S^{n-1}} \phi(r,\hat{s})\mathrm{d}\hat{s} \tag{10}$$

$$J(r) = \int_{S^{n-1}} \hat{s}\phi(r,\hat{s})\mathrm{d}\hat{s}. \tag{11}$$

By inserting the approximation (9) and similar approximations written for the source term and phase function into Eq. 4 and following the derivation in [5, 54], the P_1 approximation is obtained

$$\left(\frac{\mathrm{i}\omega}{c} + \mu_a\right)\Phi(r) + \nabla\cdot J(r) = q_0(r), \tag{12}$$

$$\left(\frac{\mathrm{i}\omega}{c} + \mu_a + \mu_s'\right)J(r) + \frac{1}{n}\nabla\Phi(r) = q_1(r), \tag{13}$$

where $\mu_s' = (1-g_1)\mu_s$ is the reduced scattering coefficient, $q_0(r)$ and $q_1(r)$ are the isotropic and dipole components of the source, and g_1 is the mean of the cosine of the scattering angle [5, 57]

$$g_1 = \int_{S^{n-1}} (\hat{s}\cdot\hat{s}')\Theta(\hat{s}\cdot\hat{s}')\mathrm{d}\hat{s}. \tag{14}$$

In the case of the Henyey–Greenstein scattering function, Eq. 5, we have $g_1 = g$.

To derive the diffusion approximation, it is further assumed that the light source is isotropic, thus $q_1(r) = 0$, and that $\frac{\mathrm{i}\omega}{c}J(r) = 0$. The latter assumption, which in time-domain case is of the form $\frac{1}{c}\frac{\partial J(r)}{\partial t} = 0$, is usually justified by specifying the condition $\mu_a \ll \mu_s'$ [5]. Utilizing these approximations, Eq. 13 gives the Fick's law

$$J(r) = -\kappa\nabla\Phi(r), \tag{15}$$

where

$$\kappa = \kappa(r) = \left(n\left(\mu_a + \mu_s'\right)\right)^{-1} \tag{16}$$

is the diffusion coefficient. Substituting Eq. 15 into Eq. 12, the frequency-domain version of the DA is obtained. It is of the form

$$-\nabla\cdot\kappa\nabla\Phi(r) + \mu_a\Phi(r) + \frac{\mathrm{i}\omega}{c}\Phi(r) = q_0(r). \tag{17}$$

The DA has an analogue in time domain as well. It is of the form

$$-\nabla\cdot\kappa\nabla\Phi(r) + \mu_a\Phi(r) + \frac{1}{c}\frac{\partial\Phi(r)}{\partial t} = q_0(r). \tag{18}$$

The time-domain and frequency-domain representations of the DA are related through Fourier transform, similarly as in the case of the RTE.

Boundary Conditions for the DA

The boundary condition (6) cannot be expressed in terms of variables of the diffusion approximation. Instead, there are a few boundary conditions that have been applied to the DA. The simplest boundary condition is the Dirichlet boundary condition which is also referred to as the zero-boundary condition. It sets the photon density to zero on the boundary; thus, $\Phi(r) = 0$, $r \in \partial\Omega$ [40, 87]. Alternatively, an extrapolated boundary condition can be used [33, 40, 87]. In the approach, the photon density is set to zero on an extrapolated boundary which is a virtual boundary outside the medium located at a certain distance from the real boundary. Both the zero-boundary condition and the extrapolated boundary condition are physically incorrect, and they have mostly been used because of their mathematical simplicity [40].

The most often used boundary condition in optical imaging is the Robin boundary condition which is also referred to as the partial current boundary condition [4, 24, 33, 40, 54, 87]. It can be derived as follows. Within the P_1 approximation framework (9), the total inward- and outward-directed photon fluxes at a point $r \in \partial\Omega$ are

$$J^-(r) = -\int_{\hat{s}\cdot\hat{n}<0} (\hat{s}\cdot\hat{n})\phi(r,\hat{s})\mathrm{d}\hat{s} = \gamma_n\Phi(r) - \frac{1}{2}\hat{v}\cdot J(r) \tag{19}$$

$$J^+(r) = \int_{\hat{s}\cdot\hat{n}>0} (\hat{s}\cdot\hat{n})\phi(r,\hat{s})\mathrm{d}\hat{s} = \gamma_n\Phi(r) + \frac{1}{2}\hat{v}\cdot J(r), \tag{20}$$

where γ_n is a dimension-dependent constant which obtains values $\gamma_2 = 1/\pi$ and $\gamma_3 = 1/4$ [57]. To derive the Robin boundary condition for the DA, it is assumed that the total inward-directed photon flux on the boundary is zero; thus,

$$J^-(r) = 0, \quad r \in \partial\Omega. \tag{21}$$

Utilizing Eq. 19 and the Fick's law (15), the Robin boundary condition can be derived. It is of the form

$$\Phi(r) + \frac{1}{2\gamma_n}\kappa\frac{\partial\Phi(r)}{\partial\hat{\nu}} = 0, \quad r \in \partial\Omega. \tag{22}$$

The boundary condition (22) can be extended to include the reflection on the boundary that is caused by different refractive indices between the object and the surrounding medium. In that case, Eq. 21 is modified to the form

$$J^-(r) = RJ^+(r), \quad r \in \partial\Omega, \tag{23}$$

where $R = R(x)$ is the reflection coefficient on the boundary $\partial\Omega$, with $0 \leq R \leq 1$ [57]. Thus, if $R = 0$, no boundary reflection occurs, and Eq. 23 is reduced into Eq. 21. The parameter R can be derived from Fresnel's law [40] or, if the refractive index of the surrounding medium is $n_{\text{out}} = 1$, by an experimental fit

$$R \approx -1.4399n_{\text{in}}^{-2} + 0.7099n_{\text{in}}^{-1} + 0.6681 + 0.0636n_{\text{in}}, \tag{24}$$

where n_{in} is the refractive index of the medium [87]. Utilizing Eqs. 19 and 20 and the Fick's law (15), the Robin boundary condition with mismatched refractive indices can be derived. It takes the form

$$\Phi(r) + \frac{1}{2\gamma_n}\kappa\zeta\frac{\partial\Phi(r)}{\partial\hat{\nu}} = 0, \quad r \in \partial\Omega, \tag{25}$$

where $\zeta = (1 + R)/(1 - R)$, with $\zeta = 1$ in the case of no surface reflection. The boundary conditions of the DA for an interface between two highly scattering materials have been discussed, for example, in [4].

The exitance, Eq. 8, can be written utilizing Eqs. 19 and 20, the Fick's law (15), and the boundary condition (25). In the DA framework, the exitance is of the form

$$\begin{aligned} J_n(r) &= J^+(r) - J^-(r) = \hat{\nu}\cdot J(r) \\ &= -\kappa\frac{\partial\Phi(r)}{\partial\hat{\nu}} = \frac{2\gamma_n}{\zeta}\Phi(r), \quad r \in \partial\Omega. \end{aligned} \tag{26}$$

Source Models for the DA

In the DA framework, light sources are usually modeled by two approximate models, namely, the collimated source model and the diffuse source model. In the case of the collimated source model, the source is modeled as an isotropic point source

$$q_0(r) = \delta(r - r_s), \tag{27}$$

where position r_s is located at a depth $1/\mu_s'$ below the source site [40, 87]. In the case of the diffuse source model, the source is modeled as an inward-directed diffuse boundary current I_s at the source position $\varepsilon_j \subset \partial\Omega$ [87]. In the case of the diffuse source model, Eq. 23 can be modified as

$$J^-(r) = RJ^+(r) + (1-R)I_s, \quad r \in \cup_j \varepsilon_j. \tag{28}$$

Then, following the similar procedure as earlier, the Robin boundary condition with the diffuse source model is obtained. It is of the form

$$\Phi(r) + \frac{1}{2\gamma_n}\kappa\zeta\frac{\partial\Phi(r)}{\partial\hat{v}} = \begin{cases} \frac{I_s}{\gamma_n}, & r \in \cup_j \varepsilon_j \\ 0, & r \in \partial\Omega \setminus \cup_j \varepsilon_j. \end{cases} \tag{29}$$

Validity of the DA

The basic condition for the validity of the DA is that the angular distribution of the radiance is almost uniform. In order to achieve this, the medium must be scattering dominated; thus, $\mu_a \ll \mu_s$. Most of the tissue types are highly scattering and the DA can be regarded as a good approximation for modeling light propagation within them. The DA has been found to describe light propagation with a good accuracy in situations in which its assumptions are valid [63, 80] and it has been successfully applied in many applications of optical tomography.

However, the condition stating that the angular distribution of the radiance must be almost uniform is violated close to the highly collimated light sources. In addition, the condition cannot be fulfilled in strongly absorbing or low-scattering tissues such as the cerebrospinal fluid which surrounds the brain and fills the brain ventricles. Furthermore, in addition to the above conditions, the DA cannot accommodate realistic boundary conditions or discontinuities at interfaces. The diffusion theory has been found to fail in situations in which its approximations are not valid such as close to the sources [35, 87] and within the low-scattering regions [37, 48].

Numerical Solution Methods for the DA

The analytical solutions of the RTE and its approximations are often restricted to certain specific geometries, and therefore, their exploitability in optical imaging is limited. Therefore, the equations describing light propagation are usually solved with numerical methods. The most often applied numerical methods are the finite difference method and the finite element method (FEM). The latter is generally regarded as more flexible when issues of implementing different boundary conditions and handling complex geometries are considered, and therefore, it is most often chosen as the method for solving equations governing light transport in tissues.

The FE model for the time-varying DA was introduced in [9]. It was later extended to address the topics of boundary conditions and source models [80, 87] and the frequency-domain case of the DA [86]. It can be regarded as the most typical approach to numerically solve the DA.

Hybrid Approaches Utilizing the DA

To overcome the limitations of the diffusion theory close to the light sources and within low-scattering and non-scattering regions, different hybrid approaches and approximate models have been developed.

The hybrid Monte Carlo diffusion method was developed to overcome the limitations of the DA close to the light sources. In the approach, Monte Carlo simulation is combined with the diffusion theory. The method was introduced in [98] to describe light reflectance in a semi-infinite turbid medium, and it was extended for turbid slabs in [97]. In the hybrid Monte Carlo diffusion approach, Monte Carlo method is used to simulate light propagation close to the light source and the DA is analytically solved elsewhere in the domain. Monte Carlo is known to describe light propagation accurately. However, it has the disadvantage of requiring a long computation time. This has effects on computation times of the hybrid Monte Carlo approaches as well. A hybrid radiative transfer–diffusion model to describe light propagation in highly scattering medium was introduced in [93]. In the approach, light propagation is modeled with the RTE close to the light sources, and the DA is used elsewhere in the domain. The solution of the RTE is used to construct a Dirichlet boundary condition for the DA on a fictitious interface within the object. Both the RTE and the DA are numerically solved with the FEM.

Different hybrid approaches and approximate models have been applied for highly scattering media with low-scattering and non-scattering regions. Methods that combine Monte Carlo simulation with diffusion theory have been applied for turbid media with low-scattering regions. The finite element approximation of the DA and the Monte Carlo simulation was combined in [41] to describe light propagation in a scattering medium with a low-scattering layer. However, also in this case, the approach suffers from the time-consuming nature of the Monte Carlo methods. Moreover, the hybrid Monte Carlo diffusion methods often require iterative mapping between the models which increases computation times even more. The radiosity–diffusion model [10, 37] can be applied for highly scattering media with non-scattering regions. The method uses the FE solution of the DA to model light propagation within highly scattering regions and the radiosity model to model light propagation within non-scattering regions. A coupled transport and diffusion model was introduced in [18]. In the model, the transport and diffusion models are coupled, and iterative mapping between the models is used for the forward solution. Furthermore, a coupled radiative transfer equation and diffusion approximation model for optical tomography was introduced in [94] and extended for domains with low-scattering regions in [92]. In the approach, the RTE is used as the forward model in sub-domains in which the assumptions of the DA are not valid and the DA is used elsewhere in the domain. The RTE and DA are coupled through boundary conditions between the RTE and DA sub-domains and solved simultaneously using the FEM.

Green's Functions and the Robin to Neumann Map

Some insight into light propagation in diffusive media can be gained by examining infinite media. In particular, verification of optical scattering and absorption parameters is frequently made with source and detector fibers immersed in a large container and far from the container walls. In a finite domain, however, we will need to use boundary conditions. We will distinguish between solutions to the homogeneous equation with inhomogeneous boundary conditions and the inhomogeneous equation with homogeneous boundary conditions. In the latter case, we can use a Green's function acting on q_0. In the former case, we use a Green's function acting on a specified boundary function.

We will use the notation G_Ω for the Green's function for the inhomogeneous form of (18) with homogeneous boundary conditions and $G_{\partial\Omega}$ for the Green's function for the homogeneous form of (18) with inhomogeneous boundary conditions, i.e., we have G_Ω solving

$$-\nabla\cdot\kappa(\mathbf{r})\nabla G_\Omega(\mathbf{r},\mathbf{r}',t,t')+\left(\mu_{\mathrm{a}}(\mathbf{r})+\frac{1}{c}\frac{\partial}{\partial t}\right)G_\Omega(\mathbf{r},\mathbf{r}',t,t')=\delta(\mathbf{r}')\delta(t') \tag{30}$$

$$\mathbf{r},\mathbf{r}'\in\Omega\backslash\partial\Omega, t>t'$$

$$G_\Omega(\mathbf{r}_{\mathrm{d}},\mathbf{r}',t,t')+2\zeta\kappa(\mathbf{r}_{\mathrm{d}})\frac{\partial G_\Omega(\mathbf{r}_{\mathrm{d}},\mathbf{r}',t,t')}{\partial\nu}=0 \tag{31}$$

$$\mathbf{r}_{\mathrm{d}}\in\partial\Omega$$

and $G_{\partial\Omega}$ solving

$$-\nabla\cdot\kappa(\mathbf{r})\nabla G_{\partial\Omega}(\mathbf{r},\mathbf{r}_{\mathrm{s}},t,t')+\left(\mu_{\mathrm{a}}(\mathbf{r})+\frac{1}{c}\frac{\partial}{\partial t}\right)G_{\partial\Omega}(\mathbf{r},\mathbf{r}_{\mathrm{s}},t,t')=0 \tag{32}$$

$$\mathbf{r}\in\Omega\backslash\partial\Omega, t>t'$$

$$G_{\partial\Omega}(\mathbf{r}_{\mathrm{d}},\mathbf{r}_{\mathrm{s}},t,t')+2\zeta\kappa(\mathbf{r}_{\mathrm{d}})\frac{\partial G_{\partial\Omega}(\mathbf{r}_{\mathrm{d}},\mathbf{r}_{\mathrm{s}},t,t')}{\partial\nu}=\delta(\mathbf{r}_{\mathrm{s}})\delta(t') \tag{33}$$

$$\mathbf{r}_{\mathrm{s}},\mathbf{r}_{\mathrm{d}}\in\partial\Omega.$$

For a given Green's function G, we define the corresponding *Green's operator* as the integral transform with G as its kernel:

$$\mathcal{G}f:=\int_{-\infty}^{\infty}\int_\Omega G(\mathbf{r},\mathbf{r}',t,t')f(\mathbf{r}',t')\mathrm{d}^{\mathrm{n}}\mathbf{r}'\mathrm{d}t.$$

For the measureable, we define the *boundary derivative operator* as

$$\mathcal{B} := -\kappa \frac{\partial}{\partial \nu},$$

where appropriate we will use the simplifying notation

$$G^{\mathcal{B}} := \mathcal{B}G$$

to mean the result of taking the boundary data for a Green's function.

Since (18) is parabolic, we must not simultaneously specify both Dirichlet and Neumann boundary conditions on the whole of $\partial\Omega$. The same is true if we convert to the frequency domain and use a complex elliptic equation to describe the propagation of the Fourier transform of Φ. Instead, we specify their linear combination through the Robin condition (25). Then, for any specified value q on $\partial\Omega$, we will get data y given by (26). The linear mapping $\Lambda q \rightarrow y$ is termed the *Robin to Neumann map* and can be considered the result of a boundary derivative operator $\mathcal{B}$ acting on the Green's operator with kernel $G_{\partial\Omega}$

$$\Lambda_{\mathrm{RtN}}(\kappa, \mu_{\mathrm{a}})q = \mathcal{B}\mathcal{G}_{\partial\Omega}q.$$

Since the Neumann data and Dirichlet data are related by (25), we may also define the Robin to Dirichlet map $\Lambda_{\mathrm{RtD}}(\kappa, \mu_{\mathrm{a}})$ and specify the relationship

$$\Lambda_{\mathrm{RtD}}(\kappa, \mu_{\mathrm{a}}) - 2\zeta\Lambda_{\mathrm{RtN}}(\kappa, \mu_{\mathrm{a}}) - I = 0 \tag{34}$$

The Forward Problem

The Robin to Neumann map is a linear operator mapping boundary sources to boundary data. For the inverse problem, we have to consider a nonlinear mapping from the space of μ_{a}, κ coefficients to the boundary data.

When considering an incoming flux J^- with corresponding boundary term q, the data is a function of one variable

$$y_q = \mathcal{F}_q \begin{pmatrix} \mu_{\mathrm{a}} \\ \kappa \end{pmatrix}, \tag{35}$$

which gives the boundary data for the particular source term $q = \mathcal{D}^-(J^-)$. Using this notation, we consider the forward mapping for a finite number of sources $\{q_j; j = 1 \ldots S\}$ as a parallel set of projections

$$\mathbf{y} = \mathcal{F} \begin{pmatrix} \mu_{\mathrm{a}} \\ \kappa \end{pmatrix}, \tag{36}$$

where

$$\mathcal{F} := (\mathcal{F}_1, \ldots, \mathcal{F}_S)^{\mathrm{T}} \tag{37}$$

$$\mathbf{y} := (y_1, \ldots, y_S)^{\mathrm{T}}. \tag{38}$$

We will consider (36) as a mapping from two continuous functions in solution space $\mu_a, \kappa \in X(\Omega) \times X(\Omega)$ to continuous functions in data space $y \in Y(\partial\Omega)$. If the data is sampled as well (which is the case in practice), then $\mathcal{F}$ is sampled at a set of measurement positions $\{\mathbf{r}_{di}; i = 1, \ldots M\}$.

The inverse problem of diffusion-based optical tomography (DOT) is to determine κ, μ_a from the values of y for all incoming boundary distributions q. If κ, μ_a are found, we can determine μ_s' through (16).

Schrödinger Form

Problem (17) can be put into Schrödinger form using the Liouville transformation. We make the change of variables $U = \kappa^{1/2}\Phi$, by which (17) becomes

$$-\kappa\nabla^2\Phi - 2\kappa^{1/2}\nabla\kappa^{1/2}\cdot\nabla\Phi + \left(\mu_a + \frac{i\omega}{c}\right)\Phi = q_0.$$

Using

$$\nabla^2 U = \kappa^{1/2}\nabla^2\Phi + 2\nabla\Phi\cdot\nabla\kappa^{1/2} + \Phi\nabla^2\kappa^{1/2}$$

leads to

$$-\nabla^2 U(\mathbf{r};\omega) + k^2(\mathbf{r};\omega)U(\mathbf{r};\omega) = \frac{q_0(\mathbf{r};\omega)}{\kappa^{1/2}(\mathbf{r})} \tag{39}$$

$$\mathbf{r} \in \Omega/\partial\Omega \tag{40}$$

$$U(\mathbf{r}_d;\omega) + 2\zeta\kappa(\mathbf{r}_d)\frac{\partial U(\mathbf{r}_d;\omega)}{\partial\nu} = \kappa^{1/2}(\mathbf{r}_d)q(\mathbf{r}_d;\omega) \tag{41}$$

$$\mathbf{r}_d \in \partial\Omega, \tag{42}$$

where

$$k^2 = \frac{\nabla^2\kappa^{1/2}}{\kappa^{1/2}} + \frac{\mu_a}{\kappa} + \frac{i\omega}{c\kappa}.$$

If k^2 is real (i.e., $\omega = 0$), there exist infinitely many κ, μ_a pairs with the same real k^2, so that the measurement of DC data cannot allow the separable unique reconstruction of κ and μ_a [6]. For $\omega \neq 0$, the unique determination of a complex k^2 should be possible by extension of the uniqueness theorem of Sylvester and Uhlmann [91]. From the complex k^2, it is in principle possible to obtain separable reconstruction of first κ from the imaginary part of k^2 and μ_a from the real part; see [74] for further discussion.

In a homogeneous medium, with constant optical parameters μ_a, κ, we can simplify (42) to

$$-\nabla^2\Phi(\mathbf{r};\omega) + k^2\Phi(\mathbf{r};\omega) = q_H(\mathbf{r};\omega), \tag{43}$$

with the same boundary condition (25) and with

$$k^2 = \left(\frac{\mu_a c + i\omega}{c\kappa}\right); \quad q_H = \frac{q_0(\mathbf{r};\omega)}{\kappa}. \tag{44}$$

This equation is also seen directly from (17) for constant κ.

The solution in simple geometries is easily derived using the appropriate Green's functions [8]. In an infinite medium, this is simply a spherical wave

$$\Phi(\mathbf{r};\omega) \equiv G(\mathbf{r},\mathbf{r}_s;\omega) = \frac{e^{\pm k|\mathbf{r}-\mathbf{r}_s|}}{|\mathbf{r}-\mathbf{r}_s|}, \tag{45}$$

where the notation $G(\mathbf{r},\mathbf{r}_s;\omega)$ defines Green's function for a source at position $\mathbf{r}_s$. Due to the real part of the wave number k, this wave is damped. This fact is the main reason that results from diffraction tomography are not always straightforwardly applicable in optical tomography. In particular, for the case $\omega = 0$, the wave is *wholly non-propagating*. Even as $\omega \to \infty$, the imaginary part of the wave number never exceeds the real part. This is a simple consequence of the parabolic nature of the diffusion approximation. Although hyperbolic approximations can be made too, they do not ameliorate the situation.

Perturbation Analysis

An important tool in scattering problems in general is the approximation of the change in field due to a change in state, developed in a series based on known functions for the reference state. There are two common approaches which we now discuss.

Born Approximation

For the Born approximation, we assume that we have a reference state $\mathbf{x}_0 = (\mu_a,\kappa)^T$, with a corresponding wave Φ, and that we want to find the *scattered* wave Φ^δ due to a change in state $\mathbf{x}^\delta = (\alpha,\beta)^T$. We have

$$\kappa = \kappa + \beta\,, \quad \mu_a = \mu_a + \alpha\,. \tag{46}$$

Note that it is not necessary to assume that the initial state is homogeneous. Putting (46) into (17) gives

$$-\nabla\cdot(\kappa+\beta)\nabla\tilde{\Phi}(\mathbf{r};\omega) + \left(\mu_a + \alpha + \frac{i\omega}{c}\right)\tilde{\Phi}(\mathbf{r};\omega) = q_0(\mathbf{r};\omega) \tag{47}$$

with

$$\tilde{\Phi} = \Phi + \Phi^\delta. \tag{48}$$

Equation 47 can be solved using the Green's operator for the reference state

$$\tilde{\Phi} = \mathcal{G}_0 \left[q_0 + \nabla \cdot \beta \nabla \tilde{\Phi} - \alpha \tilde{\Phi} \right]. \tag{49}$$

With G_0, Green's function for the reference state, we have

$$\begin{aligned} \tilde{\Phi}(\mathbf{r};\omega) &= \Phi(\mathbf{r};\omega) + \int_\Omega \left(G_0(\mathbf{r},\mathbf{r}';\omega) \nabla_{r'} \cdot \beta(\mathbf{r}') \nabla_{r'} \tilde{\Phi}(\mathbf{r}';\omega) - \alpha(\mathbf{r}') \tilde{\Phi}(\mathbf{r}';\omega) \right) \mathrm{d}^{\mathrm{n}}\mathbf{r}' \\ &= \Phi(\mathbf{r};\omega) - \int_\Omega \left(\beta(\mathbf{r}') \nabla_{r'} G_0(\mathbf{r},\mathbf{r}';\omega) \cdot \nabla_{r'} \tilde{\Phi}(\mathbf{r}';\omega) \right. \\ &\quad \left. \alpha(\mathbf{r}') G_0(\mathbf{r},\mathbf{r}';\omega) \tilde{\Phi}(\mathbf{r}';\omega) \right), \end{aligned} \tag{50}$$

where we used the divergence theorem and assumed $\beta(\mathbf{r}_\mathrm{d}) = 0; \mathbf{r}_\mathrm{d} \in \partial\Omega$.

If we define a "potential" as the differential operator

$$\mathcal{V}(\alpha,\beta) := \nabla \cdot \beta \nabla - \alpha, \tag{51}$$

we can recognize (49) as a Dyson equation and write it in the form

$$\left[\mathrm{I} - \mathcal{G}_0 \mathcal{V} \right] \tilde{\Phi} = \mathcal{G}_0 q_0. \tag{52}$$

This may by formally solved by a Neumann series,

$$\frac{\mathcal{G}_0}{[\mathrm{I} - \mathcal{G}_0 \mathcal{V}]} = \mathcal{G}_0 + \mathcal{G}_0 \mathcal{V} \mathcal{G}_0 + \mathcal{G}_0 \mathcal{V} \mathcal{G}_0 \mathcal{V} \mathcal{G}_0 + \cdots \tag{53}$$

or, equivalently, by using (48) in (50) to obtain the Born series

$$\tilde{\Phi} = \Phi^{(0)} + \Phi^{(1)} + \Phi^{(2)} + \cdots, \tag{54}$$

where

$$\begin{aligned} \Phi^{(0)} &= \Phi \\ \Phi^{(1)} &= \mathcal{G}_0 \mathcal{V} \Phi \\ \Phi^{(2)} &= \mathcal{G}_0 \mathcal{V} \mathcal{G}_0 \mathcal{V} \Phi \\ &\vdots \end{aligned}$$

Rytov Approximation

The Rytov approximation is derived by considering the logarithm of the field as a complex phase [54, 60]:

$$\Phi(\mathbf{r};\omega) = \mathrm{e}^{u(\mathbf{r};\omega)} \tag{55}$$

so that, in place of (48), we have

$$\ln \tilde{\Phi} = \ln \Phi + u^{\delta}. \tag{56}$$

Putting $\Phi = \mathrm{e}^{u_0}$ into (17), we get

$$-\Phi\nabla\cdot\kappa\nabla u_0 - \Phi\kappa\left|\nabla u_0\right|^2 + \tilde{\Phi}\left(\mu_{\mathrm{a}} + \frac{\mathrm{i}\omega}{c}\right) = q_0. \tag{57}$$

Putting $\Phi = \Phi \mathrm{e}^{u^{\delta}}$ and (46) into (17), we get

$$-\tilde{\Phi}\nabla\cdot(\kappa+\beta)\nabla\left(u_0+u^{\delta}\right) - \tilde{\Phi}(\kappa+\beta)\left|\nabla\left(u_0+u^{\delta}\right)\right|^2 + \tilde{\Phi}\left(\mu_{\mathrm{a}} + \alpha + \tfrac{\mathrm{i}\omega}{c}\right) = q_0. \tag{58}$$

Subtracting (57) from (58) and assuming $\tilde{\Phi} = \Phi$ over the support of q_0, we get

$$\begin{aligned} &-\kappa\left(2\nabla u_0\cdot\nabla u^{\delta} + \left|\nabla u^{\delta}\right|^2\right) - \nabla\cdot\kappa\nabla u^{\delta} \\ &= \nabla\cdot\beta\nabla\left(u_0+u^{\delta}\right) + \beta\left|\nabla\left(u_0+u^{\delta}\right)\right|^2 - \alpha. \end{aligned} \tag{59}$$

We now make use of the relation

$$\nabla\cdot\kappa\nabla u^{\delta}\Phi = \Phi\nabla\cdot\kappa\nabla u^{\delta} + 2\kappa\nabla\Phi\cdot\nabla u^{\delta} + u^{\delta}\nabla\cdot\kappa\nabla\Phi \tag{60}$$

$$= \Phi\left(\nabla\cdot\kappa\nabla u^{\delta} + 2\Phi\kappa\nabla u_0\cdot\nabla u^{\delta}\right) + u^{\delta}\nabla\cdot\kappa\nabla\Phi. \tag{61}$$

The last term on the right is substituted from (17) to give

$$\nabla\cdot\kappa\nabla u^{\delta} + 2\kappa\nabla u_0\cdot\nabla u^{\delta} = \frac{\nabla\cdot\kappa\nabla u^{\delta}\Phi}{\Phi} + u^{\delta}\left(\mu_{\mathrm{a}} + \frac{\mathrm{i}\omega}{c}\right) - \frac{q_0}{\Phi}. \tag{62}$$

Substituting (62) into (59) and using

$$\Phi\nabla\cdot\beta\nabla u_0 + \Phi\beta\left|\nabla u_0\right|^2 = \nabla\cdot\beta\nabla\left(\Phi u_0\right)$$

we arrive at

$$\begin{aligned} &-\nabla\cdot\kappa\nabla u^{\delta}\Phi + \left(\mu_{\mathrm{a}} + \frac{\mathrm{i}\omega}{c}\right)u^{\delta}\Phi \\ &= \nabla\cdot\beta\nabla\Phi - \alpha\Phi + \Phi\nabla\cdot\beta\nabla u^{\delta} + \kappa\left|\nabla u^{\delta}\right|^2. \end{aligned} \tag{63}$$

The approximation comes in neglecting the last two terms on the right, which are second order in the small perturbation. The left-hand side is the unperturbed

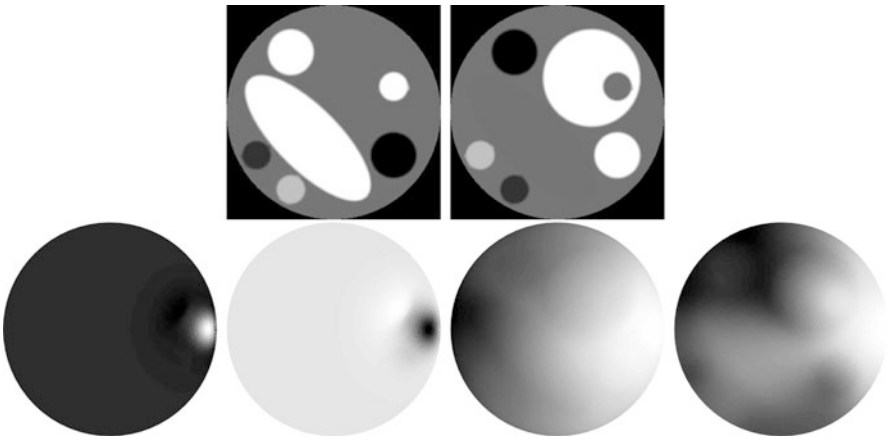


Fig. 2 *Top*: absorption and scattering images used to generate complex fields. Disk diameter 50 mm, absorption range $\mu_a \in [0.01\text{–}0.04]\,\text{mm}^{-1}$, scatter range $\mu_s' \in [1\text{–}4]\,\text{mm}^{-1}$. The complex field Φ was calculated using a 2D FEM for a δ-function source on the boundary at the 3 o'clock position. A reference field Φ was calculated for the same source and a homogeneous disk with $\mu_a = 0.025\,\text{mm}^{-1}$, $\mu_s' = 2\,\text{mm}^{-1}$. *Bottom*: the difference in fields $\Phi - \Phi$ (real and imaginary) and the difference of logs $\ln\Phi - \ln\Phi$ (real and imaginary)

operator, and so the formal solution for $u^\delta\Phi$ is again obtained through Green's operator with kernel G_0. Thus, the first-order Rytov approximation becomes

$$u^\delta(\mathbf{r};\omega) = \frac{\Phi^{(1)}(\mathbf{r};\omega)}{\Phi(\mathbf{r};\omega)} \tag{64}$$

$$= \frac{-1}{\Phi(\mathbf{r};\omega)}\big(\beta(\mathbf{r}')\nabla_{r'}G_0(\mathbf{r},\mathbf{r}';\omega)\cdot\nabla_{r'}\Phi(\mathbf{r}';\omega)$$

$$\alpha(\mathbf{r}')G_0(\mathbf{r},\mathbf{r}';\omega)\Phi(\mathbf{r}';\omega)\big)\,. \tag{65}$$

The Rytov approximation is usually argued to be applicable for larger perturbations than the Born approximation, since the neglected terms are small as long as the gradient of the field is slowly varying. See [60] for a much more detailed discussion.

Illustrations of the scattered field in the Born and Rytov formulations are shown in Fig. 2. Since in the frequency domain the field is complex, so is its logarithm. The real part corresponds to the log of the field amplitude and its imaginary part to the phase. From the images in Fig. 2, it is apparent that perturbations are more readily detected in amplitude and phase than in the field itself. This stems from the very high dynamic range of data acquired in optical tomography which in turn stems from the high attenuation and attendant damping. It is the primary motivation for the use of the Rytov approximation, despite the added complications.

Linearization

Linearization is required either to formulate a linear reconstruction problem (i.e., assuming small perturbations on a known background) or as a step in an iterative approach to the nonlinear inverse problem. We will formulate this in the frequency domain. In addition, we may work with either the wave itself, which leads to the Born approximation, or its logarithm, which leads to the Rytov approximation.

Linear Approximations

In the Born approximation to the linearized problem, we assume that the difference in measured data is given just by the first term in the Born series (54)

$$\Phi^{\delta}(\mathbf{r};\omega) \equiv \Phi^{(1)}(\mathbf{r};\omega) \tag{66}$$

$$= -\int_{\Omega} \left(\beta(\mathbf{r}')\nabla_{r'} G_{\Omega,0}(\mathbf{r},\mathbf{r}';\omega)\cdot\nabla_{r'}\Phi(\mathbf{r}';\omega)\alpha(\mathbf{r}')G_{\Omega,0}(\mathbf{r},\mathbf{r}';\omega)\Phi(\mathbf{r}';\omega)\right) \mathrm{d}^{\mathrm{n}}\mathbf{r}'.$$

From (26), we obtain for a detector at position $\mathbf{r}_{\mathrm{d}} \in \partial\Omega$

$$y(\mathbf{r}_{\mathrm{d}};\omega) = y_0(\mathbf{r}_{\mathrm{d}};\omega) + \int_{\Omega} \mathbf{K}_q^{\mathrm{T}}(\mathbf{r}_{\mathrm{d}},\mathbf{r}';\omega) \begin{pmatrix}\alpha(\mathbf{r}')\\ \beta(\mathbf{r}')\end{pmatrix} \mathrm{d}^{\mathrm{n}}\mathbf{r}', \tag{67}$$

where $\mathbf{K}_q$ is given by

$$\mathbf{K}_q(\mathbf{r}_{\mathrm{d}},\mathbf{r}';\omega) = \begin{pmatrix} G^{\mathcal{B}}_{\Omega,0}(\mathbf{r}_{\mathrm{d}},\mathbf{r}';\omega)\Phi(\mathbf{r}';\omega) \\ \nabla_{r'} G^{\mathcal{B}}_{\Omega,0}(\mathbf{r}_{\mathrm{d}},\mathbf{r}';\omega)\cdot\nabla_{r'}\Phi(\mathbf{r}';\omega)\end{pmatrix}. \tag{68}$$

The subscript q refers to the incoming flux that generates the boundary condition for the particular field Φ.

Since the Rytov approximation was derived by considering the change in the logarithm of the field, we have in place of (68) simply

$$\mathbf{K}_q^{\mathrm{Ryt}}(\mathbf{r}_{\mathrm{d}},\mathbf{r}';\omega) = \frac{1}{y_0(\mathbf{r}_{\mathrm{d}};\omega)}\mathbf{K}_q(\mathbf{r}_{\mathrm{d}},\mathbf{r}';\omega). \tag{69}$$

Assuming that we are given measured data g for a sufficient number of input fluxes, the linearized problem consists in solving for α, β from

$$y^{\delta} = \mathcal{K}_q \begin{pmatrix}\alpha\\ \beta\end{pmatrix}, \tag{70}$$

where $\mathcal{K}_q$ is a linear operator with kernel given by (68) or (69) and

$$y^{\delta} = g - y_0, \tag{71}$$

where y_0 is the data that would arise from state $\mathbf{x}_0$.

We can now distinguish between a linearized approach to the static determination of $\mathbf{x} = \mathbf{x}_0 + (\alpha, \beta)^{\mathrm{T}}$ and a *dynamic imaging* problem that assumes a reference measurement g_0. In the former case, we assume that our model is sufficiently accurate to calculate y_0. In the latter case, we use the reference measurement to solve

$$y^{\delta} = g - g_0. \tag{72}$$

This in fact is where the majority of reported results with measured data are taken. By this mechanism, inconsistencies in the modeling of the forward problem (most notably using 2D instead of 3D) are minimized. However, for static or "absolute" imaging, we still require an accurate model, even for the linearized problem.

Sensitivity Functions

If we take q to be a δ-function at a source position $\mathbf{r}_{\mathrm{s}} \in \partial\Omega$, then Φ is given by a Green's function too, and we obtain the *Photon Measurement Density Function* (PMDF)

$$\boldsymbol{\rho}(\mathbf{r}_{\mathrm{d}}, \mathbf{r}', \mathbf{r}_{\mathrm{s}}; \omega) = \begin{pmatrix} G^{\mathcal{B}}_{\Omega,0}(\mathbf{r}_{\mathrm{d}}, \mathbf{r}'; \omega) G_{\partial\Omega,0}(\mathbf{r}', \mathbf{r}_{\mathrm{s}}; \omega) \\ \nabla_{r'} G^{\mathcal{B}}_{\Omega,0}(\mathbf{r}_{\mathrm{d}}, \mathbf{r}'; \omega) \cdot \nabla_{r'} G_{\partial\Omega,0}(\mathbf{r}', \mathbf{r}_{\mathrm{s}}; \omega) \end{pmatrix} \tag{73}$$

with the Rytov form being

$$\boldsymbol{\rho}^{\mathrm{Ryt}}(\mathbf{r}_{\mathrm{d}}, \mathbf{r}', \mathbf{r}_{\mathrm{s}}; \omega) = \frac{1}{G^{\mathcal{B}}_{\partial\Omega,0}(\mathbf{r}_{\mathrm{d}}, \mathbf{r}_{\mathrm{s}}; \omega)} \boldsymbol{\rho}(\mathbf{r}_{\mathrm{d}}, \mathbf{r}', \mathbf{r}_{\mathrm{s}}; \omega).$$

It is instructive to visualize the various $\boldsymbol{\rho}$-functions which exhibit notable differences for μ_{a} and κ and between the Born and Rytov functions, as seen in Fig. 3.

Clearly

$$\mathbf{K}_q = \int_{\partial\Omega} \boldsymbol{\rho}(\mathbf{r}_{\mathrm{d}}, \mathbf{r}', \mathbf{r}_{\mathrm{s}}; \omega) q(\mathbf{r}_{\mathrm{s}}) \mathrm{d}\mathbf{r}_{\mathrm{s}} = \mathcal{G}' q,$$

where $\mathcal{G}'$ is a linear operator with kernel $\boldsymbol{\rho}$.

We can also define the linearized Robin to Neumann map

$$\Lambda'_{\mathrm{RtN}}(\mu_{\mathrm{a}}, \kappa) \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \int_{\partial\Omega} H(\mathbf{r}_{\mathrm{d}}, \mathbf{r}_{\mathrm{s}}; \omega) q(\mathbf{r}_{\mathrm{s}}) \mathrm{d}\mathbf{r}_{\mathrm{s}} = \mathcal{H} q,$$

where $\mathcal{H}$ is a linear operator with kernel H given by

$$H(\mathbf{r}_{\mathrm{d}}, \mathbf{r}_{\mathrm{s}}) = \int_{\Omega} \boldsymbol{\rho}^{\mathrm{T}}(\mathbf{r}_{\mathrm{d}}, \mathbf{r}', \mathbf{r}_{\mathrm{s}}; \omega) \begin{pmatrix} \alpha(\mathbf{r}') \\ \beta(\mathbf{r}') \end{pmatrix} \mathrm{d}^{\mathrm{n}}\mathbf{r}.$$

Fig. 3 *Top row*: sensitivity function ρ for μ_a; *left to right*: real, imaginary, amplitude, and phase; *bottom row*: the same functions for κ

Note that there are no equivalent Rytov forms. This is because the log of the Robin to Neumann map is not linear.

Adjoint Field Method

A key component in the development of a reconstruction algorithm is the use of the adjoint operator. The application of these methods in optical tomography has been discussed in detail by Natterer and coworkers [31, 74].

Taking the adjoint of $\mathcal{K}_q$ defines a mapping $\mathrm{Y}(\partial\Omega) \rightarrow \mathrm{X}(\Omega) \times \mathrm{X}(\Omega)$

$$\mathcal{K}^*{}_q b = \int_{\partial\Omega} \overline{\mathbf{K}}_q(\mathbf{r}_\mathrm{d}, \mathbf{r}'; \omega) b(\mathbf{r}_\mathrm{d}; \omega)\, \mathrm{d}S \tag{74}$$

$$= \int_{\partial\Omega} \begin{pmatrix} \overline{G}^{\mathcal{B}}_{\Omega,0}(\mathbf{r}_\mathrm{d}, \mathbf{r}'; \omega)\overline{\Phi}(\mathbf{r}'; \omega) \\ \nabla_{r'}\overline{G}^{\mathcal{B}}_{\Omega,0}(\mathbf{r}_\mathrm{d}, \mathbf{r}'; \omega) \cdot \nabla_{r'}\overline{\Phi}(\mathbf{r}'; \omega) \end{pmatrix} b(\mathbf{r}_\mathrm{d}; \omega)\, \mathrm{d}S. \tag{75}$$

Consider the reciprocity relation

$$\overline{G}^{\mathcal{B}}_{\Omega,0}(\mathbf{r}_\mathrm{d}, \mathbf{r}; \omega) = -G^*{}_{\partial\Omega,0}(\mathbf{r}, \mathbf{r}_\mathrm{d}; \omega) \tag{76}$$

with $G^*{}_{\partial\Omega,0}$, Green's function that solves the adjoint problem

$$-\nabla \cdot \kappa(\mathbf{r}) \nabla G^*{}_{\partial\Omega,0}(\mathbf{r}, \mathbf{r}_\mathrm{d}; \omega) + \left(\mu_\mathrm{a}(\mathbf{r}) - \frac{\mathrm{i}\omega}{c}\right) G^*{}_{\partial\Omega,0}(\mathbf{r}, \mathbf{r}_\mathrm{d}; \omega) = 0 \tag{77}$$

$$\mathbf{r} \in \Omega \backslash \partial\Omega$$

$$G^{*}{}_{\partial\Omega,0}(\mathbf{r},\mathbf{r}_{\mathrm{d}};\omega)+2\zeta\kappa(\mathbf{r}_{\mathrm{d}})\frac{\partial G^{*}{}_{\partial\Omega,0}(\mathbf{r},\mathbf{r}_{\mathrm{d}};\omega)}{\partial\nu}=\delta(\mathbf{r}_{\mathrm{d}};\omega) \quad \mathbf{r}_{\mathrm{d}}\in\partial\Omega. \tag{78}$$

Now, we can define a function Ψ by applying the adjoint Green's operator to the function $b\in \mathrm{Y}(\partial\Omega)$ to give

$$\Psi(\mathbf{r};\omega)=-\int_{\partial\Omega}\overline{G}^{\mathcal{B}}_{\Omega,0}(\mathbf{r}_{\mathrm{d}},\mathbf{r};\omega)b(\mathbf{r}_{\mathrm{d}};\omega)\,\mathrm{d}S \tag{79}$$

$$=\int_{\partial\Omega}G^{*}{}_{\partial\Omega,0}(\mathbf{r},\mathbf{r}_{\mathrm{d}};\omega)b(\mathbf{r}_{\mathrm{d}};\omega)\,\mathrm{d}S. \tag{80}$$

By using (78), we have that Ψ solves

$$-\nabla\cdot\kappa(\mathbf{r})\nabla\Psi(\mathbf{r};\omega)+\left(\mu_{\mathrm{a}}(\mathbf{r})-\frac{\mathrm{i}\omega}{c}\right)\Psi(\mathbf{r};\omega)=0\quad \mathbf{r}\in\Omega\backslash\partial\Omega \tag{81}$$

$$\Psi(\mathbf{r}_{\mathrm{d}};\omega)+2\zeta\kappa(\mathbf{r}_{\mathrm{d}})\frac{\partial\Psi(\mathbf{r}_{\mathrm{d}};\omega)}{\partial\nu}=b(\mathbf{r}_{\mathrm{d}};\omega)\quad \mathbf{r}_{\mathrm{d}}\in\partial\Omega \tag{82}$$

and therefore, $\mathcal{K}^{*}{}_{q}$ is given by

$$\mathcal{K}^{*}{}_{q}b=\begin{pmatrix}-\overline{\Phi}\Psi\\ -\nabla\overline{\Phi}\cdot\nabla\Psi\end{pmatrix}. \tag{83}$$

Finally, we have an adjoint form for the PMDF (73)

$$\boldsymbol{\rho}(\mathbf{r}_{\mathrm{d}},\mathbf{r}',\mathbf{r}_{\mathrm{s}};\omega)=\begin{pmatrix}-\overline{G}^{*}{}_{\partial\Omega,0}(\mathbf{r}',\mathbf{r}_{\mathrm{d}};\omega)G_{\partial\Omega,0}(\mathbf{r}',\mathbf{r}_{\mathrm{s}};\omega)\\ -\nabla_{r'}\overline{G}^{*}{}_{\partial\Omega,0}(\mathbf{r}',\mathbf{r}_{\mathrm{d}};\omega)\cdot\nabla_{r'}G_{\partial\Omega,0}(\mathbf{r}',\mathbf{r}_{\mathrm{s}};\omega)\end{pmatrix} \tag{84}$$

Time-Domain Case

In the time domain, we form the *correlation* of the propagated wave and the back-propagated residual. Equation 83 becomes

$$\mathcal{K}^{*}{}_{q}b=\begin{pmatrix}\int_0^T-\Phi(t)\Psi(t)\mathrm{d}t\\ \int_0^T-\nabla\Phi(t)\cdot\nabla\Psi(t)\mathrm{d}t\end{pmatrix}, \tag{85}$$

where $\Psi(t)$ is the solution to the adjoint equation

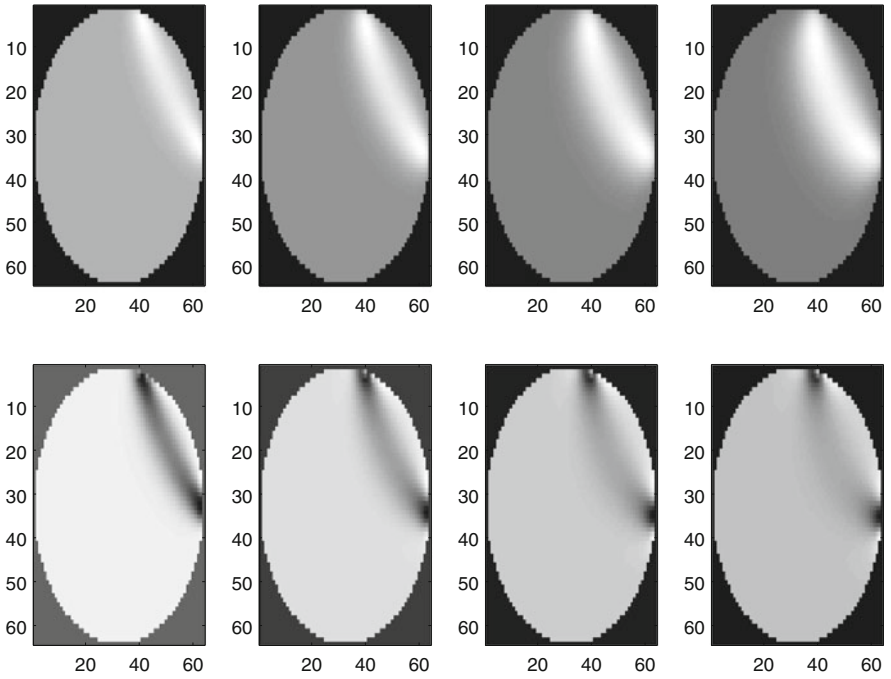


Fig. 4 *Top row*: time-window sensitivity function ρ for μ_a – *left to right*: time gates 500–1,000, 1,500–2,000, 2,500–3,000, 3,500–4,000 ps. *Bottom row*: the same functions for κ

$$\left(-\frac{1}{c}\frac{\partial}{\partial t} - \nabla\cdot\kappa(\mathbf{r})\nabla + \mu_a(\mathbf{r})\right)\Psi(\mathbf{r},t) = 0 \quad \mathbf{r}\in\Omega\backslash\partial\Omega, t\in[0,T] \tag{86}$$

$$\Psi(\mathbf{r},T) = 0, \quad \mathbf{r}\in\Omega \tag{87}$$

$$\Psi(\mathbf{r}_d,t) + 2\zeta\kappa(\mathbf{r}_d)\frac{\partial\Psi(\mathbf{r}_d,t)}{\partial\nu} = b(\mathbf{r}_d,t) \quad \mathbf{r}_d\in\partial\Omega, t\in[0,T]. \tag{88}$$

This is much more expensive, although it allows to apply temporal domain filters to optimize the effect of "early light" (that light that has undergone relatively few scattering events). In Fig. 4, the sensitivity functions over a sequence of time intervals are shown. Notice that the functions are more concentrated along the direct line of propagation for early times and become more spread out for later times.

Light Propagation and Its Probabilistic Interpretation

In time-domain systems, the source is a pulse in time which we express as a δ-function

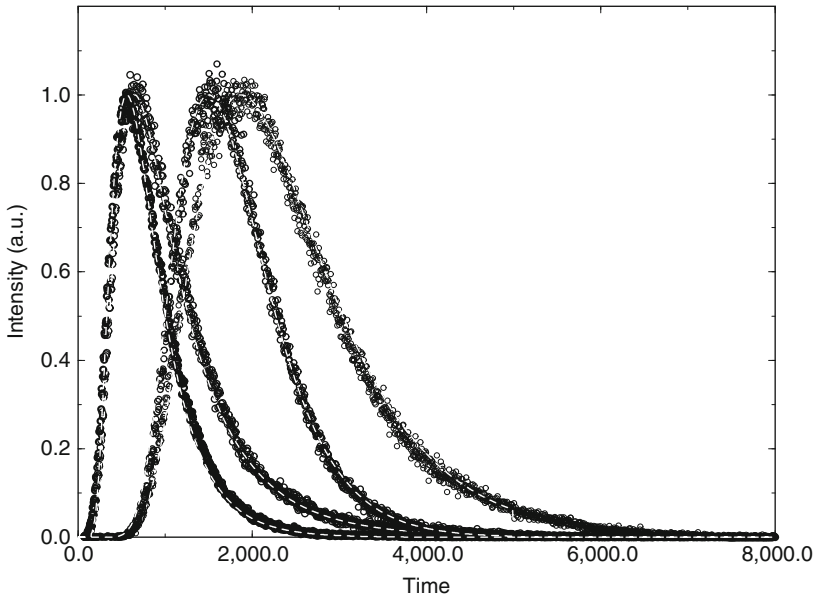


Fig. 5 Example of temporal response functions for different source-detector spacings. *Circles* represent measured data and *dashed lines* are the modeled data using a 3D finite element method. Each curve is normalized to a maximum of 1 (Data courtesy of E. Hillman and J. Hebden, University College London)

$$q(\mathbf{r}_\mathrm{d}, t) = q(\mathbf{r}_\mathrm{d})\delta(t), \tag{89}$$

where $q(\mathbf{r}_\mathrm{d})$ is the source distribution on $\partial\Omega$. Furthermore, if the input light fiber is small, the spatial distribution can be considered a δ-function too, located at a source position $\mathbf{r}_{\mathrm{s}\,j} \in \partial\Omega$

$$q(\mathbf{r}_\mathrm{d}, t) = \delta(\mathbf{r}_\mathrm{d} - \mathbf{r}_{\mathrm{s}\,j})\delta(t). \tag{90}$$

For this model, the measured signal $y(\mathbf{r}_\mathrm{d}, t)$ is the impulse response (Green's function) of the system, restricted to the boundary. When measured at a detector $\mathbf{r}_{\mathrm{d}i} \in \partial\Omega$, it is found to be a unimodal positive function of t with exponential decay to zero as $t \to \infty$. Some examples are shown in Fig. 5, showing measured data from the system described in [83] together with modeled data using a 3D finite element method. The function can be interpreted as a conditional probability density of the time of arrival of a photon, given the location of its arrival.

Consider the Green's functions for (18) in infinite space:

$$G(\mathbf{r}, \mathbf{r}', t, t') = \frac{\mathrm{e}^{-\mu_\mathrm{a} t - \frac{|\mathbf{r}-\mathbf{r}'|^2}{4\kappa(t-t')}}}{(4\pi\kappa t)^{3/2}} \quad t > t'. \tag{91}$$

Equation 91 has the form of the *Probability Density Function* (PDF) for a lossy random walk; for a fixed point in time, the distribution is spatially a Gaussian; for a fixed point in space, the distribution in time shows a sharp rise followed by an asymptotically exponential decay. In the probabilistic interpretation, we assume

$$\frac{G(\mathbf{r}, t, \mathbf{r}', t')}{\int G(\mathbf{r}, t, \mathbf{r}', t')\mathrm{d}t} \equiv P_{\mathbf{r}',t'}(t|\mathbf{r}) \tag{92}$$

as a *conditional* PDF in the sense that a photon that has arrived at a given point does so in time interval $[t, t + \delta t]$ with probability $P_{\mathbf{r}',t}(t|\mathbf{r})\delta t$. Furthermore, the absolute PDF for detecting a photon at point $\mathbf{r}$ at time t is given by

$$G(\mathbf{r}, t, \mathbf{r}', t') \equiv P_{\mathbf{r}',t'}(\mathbf{r}, t) = P_{\mathbf{r}'}(\mathbf{r})P_{\mathbf{r}',t'}(t|\mathbf{r}).$$

Here, $P_{\mathbf{r}'}(r)$ is interpreted as the relative intensity $I(\mathbf{r})/I_0$ of the detected number of photons relative to the input number.

For most PDFs based on physical phenomena, there exists a *Moment-Generating Function* (MGF)

$$M(s) = \mathbb{E}[P(t)\mathrm{e}^{st}], \tag{93}$$

where $\mathbb{E}[.]$ is the expectation operator, whence the moments (around zero) are determined by

$$m_n = \left.\frac{\partial^n M(s)}{\partial s^n}\right|_{s=0} \tag{94}$$

and in principle the PDF $P(t)$ can be reconstructed via a Taylor series for its MGF

$$M(s) = m_0 + m_1 s + \cdots m_n \frac{s^n}{n!} + \tag{95}$$

However, explicit evaluation of this series is impractical. Furthermore, we may assume that only a small number of independent moments exist, in which case we reconstruct the series implicitly given only the first few moments. In the results presented here, only the first three moments m_0, m_1, and m_2 are used. They have the physical interpretations

$$\begin{array}{ll} m_0 & \text{Total intensity } I(\mathbf{r}) \\ \dfrac{m_1}{m_0} & \text{Mean time } \quad < t > (\mathbf{r}) \\ \dfrac{m_2}{m_0} - \left(\dfrac{m_1}{m_0}\right)^2 & \text{Variance time } \sigma_t^2(\mathbf{r}). \end{array}$$

In order to test the validity of the moment method to construct the time-varying solution, we created a finite element model of a $4 \times 7 \times 4$ cm slab. The optical

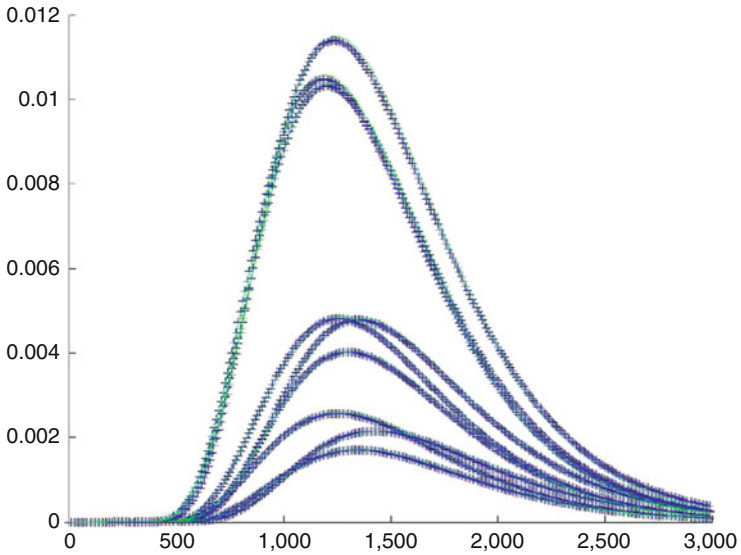


Fig. 6 An example of the output time-of-flight histograms at each of nine detectors on the transmission surface of a slab. *Solid curves* are computed from (95) using the zeroth, first, and second moments and implicit extrapolation; *crosses* are computed using finite differencing in time of the system (18) and 300 steps of size 10 ps

parameters were set to an arbitrary heterogeneous distribution by adding a number of Gaussian blobs of different amplitude and spatial width in both μ_a and κ to a background of $\mu_a = 0.1\ \text{cm}^{-1}, \kappa = 0.03\ \text{cm}$. A source was placed at the center of one face of the slab and nine detectors placed in a rectangular array on the opposite face of the slab. The time-of-flight histogram of transmitted photons at each detector was calculated in two ways: (1) by solving the time-dependent system (18) using a fully implicit finite differencing step in time (2) by solving for the moments m_0, m_1, m_2 using (94) and deriving the time-varying solution via (95).

One case is shown in Fig. 6. The comparison is virtually perfect despite the grossly heterogeneous nature of the example which precludes the exact specification of a Green's function. The moment-based method is several hundred times faster.

4 Numerical Methods and Case Examples

Image Reconstruction in Optical Tomography

Optical tomography is generally recognized as a nonlinear inverse problem; linear methods can certainly be applied (e.g., [85]) but are limited to the case of small perturbations on a known background.

The following is an overview of the general approach: we construct a physically accurate model that describes the progress of photons from the source optodes

through the media and to the detector optodes. This is termed the *forward problem* (section "The Forward Problem"). This model is parameterized by the spatial distribution of scattering and absorption properties in the media. We adjust these properties iteratively until the predicted measurements from the forward problem match the physical measurements from the device. This is termed the *inverse problem*.

The model-based approach is predicated on the implicit assumption that there exists in principle an "exact model" given by the physical description of the problem, and the task is to develop a computational technique that matches measured data within an accuracy below the expected level of measurement error. In other words, we assume that model inaccuracies are insignificant with respect to experimental errors. However, the computational effort in constructing a forward model of sufficient accuracy can be prohibitive at best. In addition, the physical model may have shortcomings that lead to the data being outside the range of the forward model (nonexistence of solution).

In the *approximation error method* [57], we abandon the need to produce an "exact model." Instead, we attempt to determine the statistical properties of the modeling errors and compensate for them by incorporating them into the image reconstruction using a Bayesian approach (section "Approximation Error Approach"). The steps involved in using the approximation error theory are (1) the construction and sampling of a prior, (2) the construction of a mapping between coarse and fine models, (3) calculation of forward data from the samples on both coarse and fine models, and (4) statistical estimation of the mean and covariance of the differences between data from the coarse and fine models. In [11], this technique was shown to result in reconstructed images using a relatively inaccurate forward model that were of almost equal quality to those using a more accurate forward model; the increase in computational efficiency was an order of magnitude.

Reconstruction from optical measurements is difficult because it is fundamentally ill posed. We usually aim to reconstruct a larger number of voxels than there are measurements (which results in *nonuniqueness*), and the presence of noisy measurements can result in an exponential growth in the image artifacts. In order to stabilize the reconstruction, *regularization* is required. Whereas such regularization is often based on ad hoc considerations for improving image quality, the Bayesian approach provides a rigorous framework in which the reconstructed images are chosen to belong to a distribution with principled characteristics (the prior).

Bayesian Framework for Inverse Optical Tomography Problem

In the Bayesian framework for inverse problems, all unknowns are treated and modeled as random variables [57]. The measurements are often considered as random variables also in non-Bayesian framework for inverse problems. The actual modeling of the measurements as random variables is, however, often implicit, which is most manifest when least squares functionals are involved in the formulation of the problem. In the Bayesian framework, however, both the measurements

and the unknowns are *explicitly* modeled as random variables. The construction of the *likelihood (observation) models* and the *prior models* is the starting point of the Bayesian approach to (inverse) problems.

Once the probabilistic models for the unknowns and the measurement process have been constructed, the *posterior distribution* $\pi(x \mid y)$ is formed, which distribution reflects the uncertainty of the interesting unknowns x given the measurements y. This distribution can then be explored to answer all questions which can be expressed in terms of probabilities. For general discussion of Bayesian inference, see, for example, [21].

Bayesian inverse problems are a special class of problems in Bayesian inference. Usually, the dimension of a feasible representation of the unknowns is significantly larger than the number of measurements, and thus, for example, a maximum likelihood estimate is either impossible or extremely unstable to compute. In addition to the instability, the variances of the likelihood model are almost invariably much smaller than the variances of the prior models. The posterior distribution is often extremely narrow and, in addition, may be a nonlinear manifold.

Bayesian Formulation for the Inverse Problem

In the following, we denote the unknowns with the vector x, the measurements with y, and all probability distributions (densities) by π. Typically, we would have $x = (\mu_a, \mu_s)$, with μ_a and μ_s identified with the coordinates in the used representations.

The complete statistical information of all the random variables is given by the joint distribution $\pi(x, y)$. This distribution expresses all the uncertainty of the random variables. Once the measurements y have been obtained, the uncertainty in the unknowns x is (usually) reduced. The measurements are now reduced from random variables to numbers, and the uncertainty of x is expressed as the posterior distribution $\pi(x \mid y)$. This distribution contains all information on the uncertainty of the unknowns x when the information on measurements y is utilized.

The conditional distribution of the measurements given the unknown is called the *likelihood distribution* and is denoted by $\pi(y \mid x)$. The marginal distribution of the unknown is called the *prior (distribution)* and is denoted by $\pi(x)$. By the definition of conditional probability, we have

$$\pi(x, y) = \pi(y \mid x)\pi(x) = \pi(x \mid y)\pi(y). \tag{96}$$

Furthermore, the marginal distributions can be obtained by marginalizing (integrating) over the remaining variables, that is, $\pi(x) = \int \pi(x, y)\, \mathrm{d}y$ and $\pi(y) = \int \pi(x, y)\, \mathrm{d}x$. The following rearrangement is called Bayes' theorem

$$\pi(x \mid y) = \pi(y)^{-1}\pi(y \mid x)\pi(x). \tag{97}$$

If we were given the joint distribution, we could simply use the above definitions to compute the posterior distribution. Unfortunately, the joint distribution is practically never available in the first place. However, it turns out that in many cases,

the derivation of the likelihood density is a straightforward task. Also, a feasible probabilistic model for the unknown can often be obtained. Then, one can use Bayes' theorem to obtain the posterior distribution. The demarcation between the Bayesian and frequentist paradigms is that, here, the posterior is obtained by using a (prior) model for the distribution of the unknown rather than the marginal density, which cannot be computed since the joint distribution is not available in the first place. We stress that all distributions have to be interpreted as models.

Inference

Point estimates are the Bayesian counterpart of the "solutions" suggested by regularization methods. The most common point estimates are the *maximum a posteriori* estimate (MAP) and the *conditional mean* estimate (CM). Let the unknowns and measurements be the finite-dimensional random vectors $x \in \mathbb{R}^N$, $y \in \mathbb{R}^M$.

The computation of the MAP estimate is an optimization problem, while the computation of the CM estimate is an integration problem:

$$x_{\mathrm{MAP}} = \operatorname{sol} \max_x \pi(x \mid y) \tag{98}$$

$$x_{\mathrm{CM}} = \mathbb{E}(x \mid y) = \int x\, \pi(x \mid y)\, \underset{\cdot}{\mathrm{x}}, \tag{99}$$

where sol reads as "solution of" the maximization problem, $\mathbb{E}(\cdot)$ denotes expectation, and the integral in (99) is an N-tuple integral.

The most common estimate of spread is the *conditional covariance*

$$\Gamma_{x|y} = \int (x - \mathbb{E}(x \mid y))(x - \mathbb{E}(x \mid y))^{\mathrm{T}}\, \pi(x \mid y)\, \underset{\cdot}{\mathrm{x}}. \tag{100}$$

Here, $\Gamma_{x \mid y}$ is an $N \times N$ matrix and the integral (100) refers to a matrix of associated integrals.

Often, the marginal distributions of single variables are also of interest. These are formally obtained by integrating over all other variables

$$\pi(x_\ell \mid y) = \int_{x_{-\ell}} \pi(x \mid y)\, \underset{\cdot}{\mathrm{x}}_{-\ell}, \tag{101}$$

where the notation $(\cdot)_{-\ell}$ refers to all components *excluding* the ℓth component. Note that $\pi(x_\ell \mid y)$ is a function of a single variable and can be visualized by plotting. The *credibility intervals* are the Bayesian counterpart to the frequentist confidence intervals, but the interpretation is different. The $p\%$-credibility interval is a subset which contains $p\%$ of the probability mass of the *posterior distribution*.

Likelihood and Prior Models

The likelihood model $\pi(y \mid x)$ consists of modeling the forward problems and the related observational errors. In the likelihood model, all unknowns are treated as fixed. The most common likelihood model is based on the additive error model

$$y = \mathcal{F}(x) + e,$$

where e is the additive error term with distribution $\pi_e(e)$ which is usually modeled as mutually independent with x. In this case, we can get rid of the unknown additive error term by pre-marginalizing over it. We have formally $\pi(y \mid x, e) = \delta(y - \mathcal{F}(x) + e)$ and using the Bayes theorem

$$\pi(y \mid x) = \pi_e(y - \mathcal{F}(x)).$$

For more general derivation and other likelihood models, see, for example, [57].

In the special case of Gaussian additive errors $\pi_e(e) = \mathcal{N}(e_*, \Gamma_e)$, we get

$$\pi(y \mid x) \propto \exp\left(-\frac{1}{2}\|L_e(y - \mathcal{F}(x) - e_*)\|^2\right),$$

where $\Gamma_e^{-1} = L_e^{\mathrm{T}} L_e$. In the very special case of $\pi_e(e) = \mathcal{N}(0, \gamma^2 I)$, we of course get the ordinary least squares functional for the posterior potential. This particular model, however, should always be subjected to assessment since it usually corresponds to an idealized measurement system.

For prior models $\pi(x)$ for the unknowns whose physical interpretation is a distributed parameter, Markov random fields are a common choice. The most common type is an improper prior model of the form

$$\pi(x) \propto \exp\left(-\frac{1}{2}\|L_x(x - x_*)\|^2\right), \tag{102}$$

where L_x is derived from a spatial differential operator. For example, $\|L_x(x - x_*)\|^2$ might be a discrete approximation for $\int_\Omega |\Delta x(\vec{r})|^2 \, \mathrm{d}\vec{r}$. Such improper prior models may work well technically since the null space of L_x is usually such that it is annihilated in the posterior model. It must be noted, however, that there are constructions that yield proper prior models [57, 59, 67]. These are needed, for example, for the construction of approximation error models discussed in section "Approximation Error Approach." Moreover, structural information related to inhomogeneities and anisotropicity of smoothness can be decoded in these models [59].

Nonstationary Problems

Inverse problems in which the unknowns are time varying are referred to as *nonstationary inverse problems* [57]. These problems are also naturally cast in the Bayesian framework. Nonstationary inverse problems are usually written as evolution–observation models in which the evolution of the unknown is typically modeled as a stochastic process. The related algorithms are sequential and in the most general form are of the Markov chain Monte Carlo type [32]. However, the most commonly used algorithms are based on the Kalman recursions [3, 57, 61].

A suitable statistical framework for dealing with unknowns that are modeled with stochastic processes and which are observed either directly or indirectly is the *state*

estimation framework. In this formalism, the unknown is referred to as *the state variable*, or simply *the state*. For treatises on state estimation and Kalman filtering theory in general, see, for example, [3, 34]. For the general nonlinear non-Gaussian treatment, see [32], and for state estimation with inverse problems, see [57].

The general discrete time *state space representation* of a dynamical system is of the form

$$x_{k+1} = F_k(x_k, w_k) \tag{103}$$

$$y_k = A_k(x_k, v_k), \tag{104}$$

where w_k is the *state noise process*, v_k is the *observation noise process*, and (103) and (104) are the *evolution model* and *observation model*, respectively. Here, the evolution model replaces the prior model in stationary inverse problems, while the observation model is usually the same as the (stationary) likelihood model. We do not state the exact assumptions here, since the assumptions may vary somewhat resulting in different variations of Kalman recursions; see, for example, [3]. It suffices here to state that the sequences of mappings F_t and A_t are assumed to be known and that the state and observation noise processes are temporally uncorrelated and that their (second-order, possibly time-varying) statistics are known. With these assumptions, the state process is a first-order Markov process. The first-order Markov property facilitates recursive algorithms for the state estimation problem. The Kalman recursions were first derived in [61].

Formally, the state estimation problem is to compute the distribution of a state variable $x_k \in \mathbb{R}^N$ given a set of observations $y_j \in \mathbb{R}^M$, $j \in \mathcal{I}$ where $\mathcal{I}$ is a set of time indices. In particular, the aim is to compute the related conditional means and covariances. Usually, $\mathcal{I}$ is a contiguous set of indices and we denote $Y_\ell = (y_1, \ldots, y_\ell)$.

We can then state the following common state estimation problems:

- *Prediction.* Compute the conditional distribution of x_k given Y_ℓ, $k > \ell$.
- *Filtering.* Compute the conditional distribution of x_k given Y_ℓ, $k = \ell$.
- *Smoothing.* Compute the conditional distribution of x_k given Y_ℓ, $k < \ell$.

The solution of the state estimation problems in linear Gaussian cases is usually carried out by employing the Kalman filtering or smoothing algorithms that are based on Kalman filtering. These are recursive algorithms and may be either real-time, online, or batch-type algorithms.

In nonlinear and/or non-Gaussian cases, extended Kalman filtering (EKF) variants are usually employed. The EKF algorithms form a family of estimators that do not possess any optimality properties. For many problems, however, the EKF algorithms provide feasible state estimates. For EKF algorithms, see, for example, [3, 57]. Since the observation models with optical tomography are nonlinear, the EKF algorithms are a natural choice for nonstationary DOT problems; see [30, 66, 81].

The idea in extended Kalman filters is straightforward: the nonlinear mappings are approximated with the affine mappings given by the first two terms of the Taylor expansion. The version of extended Kalman filter that is most commonly used is the *local linearization version*, in which version the mappings are linearized at the best currently available state estimates, either the predicted or the filtered state. This necessitates the recomputation of the Jacobians $\partial A_t / \partial x_t$ at each time instant.

The EKF recursions take the form

$$x_{k|k-1} = F_{k-1}(x_{k-1|k-1}) + s_{k-1} + B_{k-1}(u_{k-1}) \tag{105}$$

$$\Gamma_{k|k-1} = J_{F_{k-1}} \Gamma_{k-1|k-1} J_{F_{k-1}}{}^{\mathrm{T}} + \Gamma_{w_{k-1}} \tag{106}$$

$$K_k = \Gamma_{k|k-1} J_{A_k}{}^{\mathrm{T}} \left(J_{A_k} \Gamma_{k|k-1} J_{A_k}{}^{\mathrm{T}} + \Gamma_{v_k} \right)^{-1} \tag{107}$$

$$\Gamma_{k|k} = \left(I - K_k J_{A_k} \right) \Gamma_{k|k-1} \tag{108}$$

$$x_{k|k} = x_{k|k-1} + K_k \left(y_k - A_k(x_{k|k-1}) \right), \tag{109}$$

where $x_{k|k-1}$ and $x_{k|k}$ are the prediction and filtering estimates, respectively, and $\Gamma_{k|k-1}$ and $\Gamma_{k|k}$ are the approximate prediction and filtering covariances, respectively. Note that the Jacobian mappings (linearizations) are needed only in the computation of the covariances and the Kalman gain K_t.

The applications of EKF algorithms to nonstationary DOT problems have been considered in [30, 66, 81]. In [66], a random walk evolution model was constructed and used for tracking of targets in a cylindrical tank geometry. In [81], a cortical mapping problem was considered, in which the evolution model was augmented to include auxiliary periodic processes to allow for separation of cyclical phenomena from evoked responses. In [30], an elaborate physiological model was added to that of [81] to form the evolution model.

Approximation Error Approach

The approximation error approach was introduced in [57, 58] originally to handle pure model reduction errors. For example, in electrical impedance (resistance) tomography (EIT, ERT) and deconvolution problems, it was shown that significant model reduction is possible without essentially sacrificing the quality of estimates. With model reduction, we mean that very low-dimensional finite element approximations can be used for the forward problem. The approximation error approach relies heavily on the Bayesian framework of inverse problems, since the approximation and modeling errors are modeled as additive errors *over the prior model*.

In this following, we discuss the approximation error approach in a setting in which one distributed parameter is of interest, while another one is not, and there are additional uncertainties that are related, for example, to unknown boundary data. In addition, we formulate the problem to take into account model reduction errors. In the case of optical tomography, this would mean using very approximate forward solvers, for example.

Let now the unknowns be (μ_a, μ_s, ξ, e), where e represents additive errors and ξ represents auxiliary uncertainties, such as unknown boundary data, and μ_a is of interest only. Let

$$y = \bar{A}(\mu_a, \mu_s, \xi) + e \in \mathbb{R}^m$$

denote an accurate model for the relation between the measurements and the unknowns.

In the approximation error approach, we proceed as follows. Instead of using the accurate forward model $(\mu_a, \mu_s, \xi) \mapsto \bar{A}(\mu_a, \mu_s, \xi)$ with (μ_a, μ_s, ξ) as the unknowns, we fix the random variables $(\mu_s, \xi) \leftarrow (\mu_{s,0}, \xi_0)$ and use a computationally (possibly drastically reduced) approximate model

$$\mu_a \mapsto A(\mu_a, \mu_{s,0}, \xi_0).$$

Thus, we write the measurement model in the form

$$y = \bar{A}(\mu_a, \mu_s, \xi) + e \tag{110}$$

$$= A(\mu_a, \mu_{s,0}, \xi_0) + \big(\bar{A}(\mu_a, \mu_s, \xi) - A(\mu_a, \mu_{s,0}, \xi_0)\big) + e \tag{111}$$

$$= A(\mu_a, \mu_{s,0}, \xi_0) + \varepsilon + e, \tag{112}$$

where we define the *approximation error* $\varepsilon = \varphi(\mu_a, \mu_s, \xi) = \bar{A}(\mu_a, \mu_s, \xi) - A(\mu_a, \mu_{s,0}, \xi_0)$. Thus, the approximation error is the discrepancy of predictions of the measurements (given the unknowns) when using the accurate model $\bar{A}(\mu_a, \mu_s, \xi)$ and the approximate model $A(\mu_a, \mu_{s,0}, \xi_0)$.

Using the Bayes' formula repeatedly, it can be shown that

$$\pi(y \mid x) = \int \pi_e(y - A(x, \mu_{s,0}, \xi_0) - \varepsilon)\pi_{\varepsilon|x}(\varepsilon \mid x)\, d\varepsilon \tag{113}$$

since e and x are mutually independent. Note that (112) and (113) are exact.

In the approximation error approach, the following Gaussian approximations are used: $\pi_e \approx \mathcal{N}(e_*, \Gamma_e)$ and $\pi_{\varepsilon|x} \approx \mathcal{N}(\varepsilon_{*,\mu_a}, \Gamma_{\varepsilon|\mu_a})$. Let the normal approximation for the joint density $\pi(\varepsilon, \mu_a)$ be

$$\pi(\varepsilon, \mu_a) \propto \exp\left\{-\frac{1}{2}\begin{pmatrix} \varepsilon - \varepsilon_* \\ \mu_a - \mu_{a,*} \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} \Gamma_{\varepsilon\varepsilon} & \Gamma_{\varepsilon\mu_a} \\ \Gamma_{\mu_a\varepsilon} & \Gamma_{\mu_a\mu_a} \end{pmatrix}^{-1} \begin{pmatrix} \varepsilon - \varepsilon_* \\ \mu_a - \mu_{a,*} \end{pmatrix}\right\} \tag{114}$$

whence

$$\varepsilon_{*,\mu_a} = \varepsilon_* + \Gamma_{\varepsilon\mu_a}\Gamma_{\mu_a\mu_a}^{-1}(\mu_a - \mu_{a,*}) \tag{115}$$

$$\Gamma_{\varepsilon|\mu_a} = \Gamma_{\varepsilon\varepsilon} - \Gamma_{\varepsilon\mu_a}\Gamma_{\mu_a\mu_a}^{-1}\Gamma_{\mu_a\varepsilon}. \tag{116}$$

Define the normal random variable $\nu = e + \varepsilon$ so that $\nu \mid \mu_a \sim \mathcal{N}(\nu_{*|\mu_a}, \Gamma_{\nu|\mu_a})$. Thus, we obtain for the approximate likelihood distribution

$$\pi(y \mid \mu_a) \approx \mathcal{N}(y - A(\mu_a, \mu_{s,0}, \xi_0) - v_{*|\mu_a}, \Gamma_{v|\mu_a}).$$

Since we are after computational efficiency, a normal approximation $\pi(\mu_a) \approx \mathcal{N}(\mu_{a,*}, \Gamma_{\mu_a})$ for the prior model is also usually employed. Thus, we obtain the approximation for the posterior distribution

$$\pi(\mu_a \mid y) \propto \pi(y \mid \mu_a)\pi(\mu_a) \tag{117}$$

$$\propto \exp\left(-\frac{1}{2}\|L_{v|\mu_a}(y - A(\mu_a, \mu_{s,*}, \xi_*) - v_{*|\mu_a})\|^2 \right. \tag{118}$$

$$\left. + \|L_{\mu_a}(\mu_a - \mu_{a,*})\|^2\right), \tag{119}$$

where $\Gamma^{-1}_{v\,|\,\mu_a} = L^{\mathrm{T}}_{v|\mu_a} L_{v|\mu_a}$ and $\Gamma^{-1}_{\mu_a\mu_a} = L^{\mathrm{T}}_{\mu_a} L_{\mu_a}$. See [68] or more details on the particular problem of marginalizing over the scattering coefficient.

The approximation error approach has been applied to various kinds of approximation and modeling errors as well as other inverse problems. Model reduction, domain truncation, and unknown anisotropy structures in diffuse optical tomography were treated in [12, 45, 46, 67]. Missing boundary data in the case of image processing and geophysical EIT were considered in [23] and [70], respectively. Furthermore, in [76–78], the problem of recovery from simultaneous geometry errors and model reduction was found to be possible. In [95], the radiative transfer model was replaced with the diffusion approximation. It was found that also in this kind of a case, the statistical structure of the approximation errors enabled the use of a significantly less complex model, again simultaneously with significant model reduction for the diffusion approximation. But also here, both the absorption and scattering coefficients were estimated simultaneously.

The approximation error approach was extended to nonstationary inverse problems in [51] in which linear nonstationary (heat transfer) problems were considered and in [50] and [52] in which nonlinear problems and state space identification problems were considered, respectively. A modification in which the approximation error statistics can be updated with accumulating information was proposed in [53] and an application to hydrogeophysical monitoring in [71].

Experimental Results

In this section, we show an example where the error model is employed for compensating the modeling errors caused by reduced discretization accuracy h and experimental DOT data is used for the observations.

Experiment and Measurement Parameters

The experiment was carried out with the frequency-domain (FD) DOT instrument at Helsinki University of Technology [75]. The measurement domain Ω is a cylinder with radius $r = 35$ mm and height 110 mm; see Fig. 7. The target consists of homogeneous material with two small cylindrical perturbations, as illustrated in

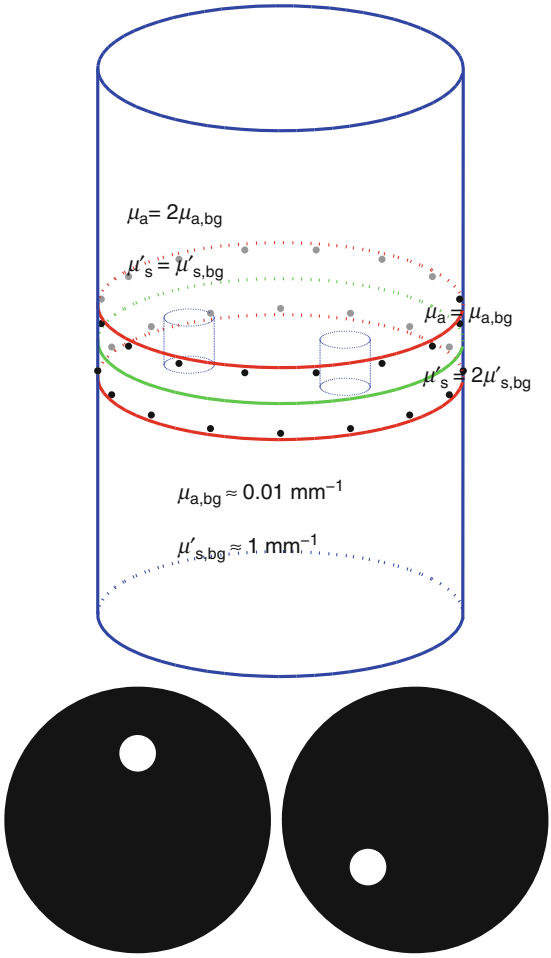


Fig. 7 *Top*: Measurement domain Ω. The *dots* denote the location of the sources and detectors. The (*red*) lines above and below the sources and detectors denote the truncated model domain Ω. The *green line* denotes the central slice of the domain Ω. *Bottom*: Central slice of the target (μ_a *left*, μ'_s *right*)

Fig. 7. The background optical properties of the phantom are approximately $\mu_{a,bg} = 0.01\,mm^{-1}$ and $\mu_{s,bg} = 1\,mm^{-1}$ at wavelength $\lambda \approx 800$ nm. The cylindrical perturbations, which both have diameter and height of 9.5 mm, are located such that the central plane of the perturbations coincides with the central xy-plane of the cylinder domain Ω. For an illustration of the cross sections of μ_a and μ'_s, see bottom row in Fig. 7. The optical properties of perturbation 1 are approximately $\mu_{a,p1} = 0.02\,mm^{-1}$, $\mu_{s,p1} = 1\,mm^{-1}$ (i.e., purely absorption contrast) and the properties of perturbation 2 are $\mu_{a,p2} = 0.01\,mm^{-1}$, $\mu_{s,p2} = 2\,mm^{-1}$ (i.e., purely scatter contrast), respectively. The source and detector configuration in the experiment consisted of 16 sources and 16 detectors arranged in interleaved order on two rings located 6 mm above and below the central xy-plane of the cylinder domain. The locations of sources and detectors are shown with dots in Fig. 7. The measurements were carried out at $\lambda = 785$ nm with an optical power of 8 mW and

modulation frequency $2\pi f = 100\,\text{MHz}$. The log amplitude and phase shift of the transmitted light was recorded at 12 farthermost detector locations for each source, leading to a real-valued measurement vector

$$y = \begin{pmatrix} \text{re}(\log(z)) \\ \text{im}(\log(z)) \end{pmatrix} \in \mathbb{R}^{384}$$

for the experiment. The statistics of measurement noise in the measurement y are not known. Thus, we employ the same implicit (ad hoc) noise model that was used for reconstructions from the same measurement realization in [88]. The noise model is

$$e \sim \mathcal{N}(0, \Gamma_e),$$

where Γ_e is a diagonal data scaling matrix which is tuned such that the initial (weighted) least squares (LS) residual

$$\|L_e(y - y_0)\|^2 = 2, \quad \Gamma_e^{-1} = L_e^{\mathrm{T}} L_e$$

between the measured data y and forward solution y_0 at the initial guess $x = x_0$ becomes unity for both data types (log amplitude re(log(z)) and phase im(log(z))).

Prior Model

In this study, we use a proper Gaussian smoothness prior as the prior model for the unknowns. The absorption and scatter images μ_a and μ_s' are modeled as mutually independent Gaussian random fields with a joint prior model

$$\pi(x_\delta) \propto \exp\left\{-\frac{1}{2}\|L_{x_\delta}(x_\delta - x_{\delta *})\|^2\right\}, \quad L_{x_\delta}^{\mathrm{T}} L_{x_\delta} = \Gamma_{x_\delta}^{-1}, \tag{120}$$

where

$$\Gamma_{x_\delta} = \begin{pmatrix} \Gamma_{\mu_\text{a}} & 0 \\ 0 & \Gamma_{\mu_\text{s}'} \end{pmatrix}.$$

The construction of the blocks Γ_{μ_a} and $\Gamma_{\mu_\text{s}'}$ has been explained for a two-dimensional case in [11], and the extension to three-dimensional case is straightforward. The parameters in the prior model were selected as follows. The correlation length for both μ_a and μ_s' in the prior was set as 11 mm. The correlation length can be viewed (roughly) as our prior estimate about the expected spatial size of perturbations in the target domain. The prior mean for absorption and scatter were set as $\mu_{\text{a}*} = 0.01\,\text{mm}^{-1}$ and $\mu_{\text{s}*} = 1\,\text{mm}^{-1}$, and the marginal standard deviations of absorption and scatter in each voxel were chosen such that $3\sigma_{\mu_\text{a}} = 0.01$ and $3\sigma_{\mu_\text{s}'} = 1$, respectively. This choice corresponds to assuming that the values of absorption and scatter are expected to lie within the intervals $\mu_\text{a} \in [0, 0.02]$ and

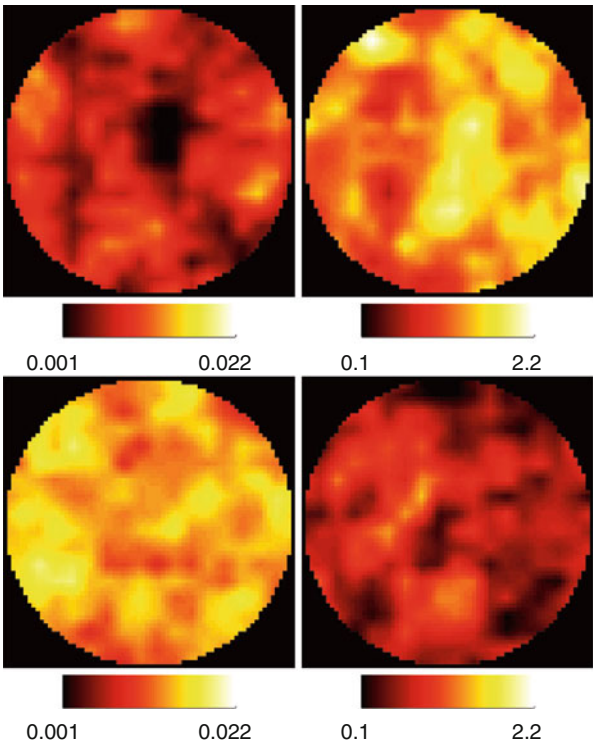


Fig. 8 Two random samples from the prior density (120). The images display the cross section of the 3D parameters at the central slice of the cylinder domain Ω. *Left*: Absorption μ_a. *Right*: Scatter μ_s'

$\mu_s' \in [0,\ 2]$ with *prior* probability of 99.7 %. Figure 8 shows two random samples from the prior model.

Selection of FEM Meshes and Discretization Accuracy

To select the discretization accuracy δ for the accurate forward model $A_{\Omega,\delta}(x_\delta)$, we adopted a similar procedure as in [11]. In this process, we computed relative error in the FEM solution with respect to the discretization level h and identified δ as that mesh density beyond which the relative error

$$\frac{\|A_{\Omega,h} - A_{\Omega,h'}\|}{\|A_{\Omega,h'}\|}$$

in both amplitude and phase parts of the forward solution was stabilized. The mesh for the reference model $A_{\Omega,h'}$ in the convergence analysis consisted of $N_n = 253{,}981$ node points and $N_e = 1{,}458{,}000$ (approximately) uniform tetrahedral elements. We found that the errors in the FEM solution were stabilized when using a (uniform) tetrahedral mesh with (approximately) 150,000 nodes or more, and thus,

Table 1 Mesh details for test case. N_n is the number of nodes, N_e is the number of tetrahedral elements in the mesh, and n_p is the number of voxels in the representation of μ_a and μ_s'. t is the wall clock time for one forward solution

Model	N_n	N_e	n_p	t (s)
$A_{\Omega,\delta}$	148,276	843,750	7,668	178
$A_{\Omega,h}$	2,413	11,664	7,668	0.4

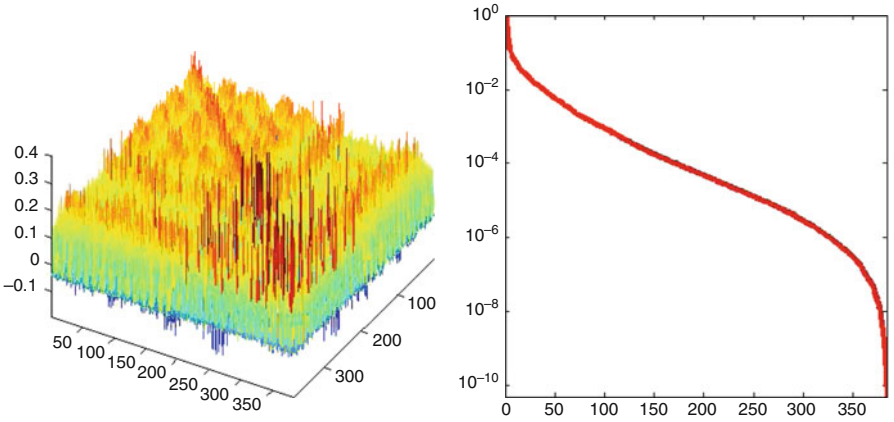


Fig. 9 Modeling error between the accurate model $A_{\Omega,\delta}$ and target model $A_{\Omega,h}$; see Table 1. *Left*: Covariance structure of the approximation error ε. The displayed quantity is the signed standard deviation $\text{sign}(\Gamma_\varepsilon) \cdot \sqrt{|\Gamma_\varepsilon|}$, where the product refers to the element-by-element (array) multiplication. *Right*: Normalized eigenvalues $\lambda/\lambda_{\max}$ of Γ_ε. of Γ_ε

we chose for the accurate model $A_{\Omega,\delta}$ a mesh with $N_n = 148{,}276$ node points. For the target model $A_{\Omega,h}$, we chose a mesh with $N_n = 2{,}413$ nodes; see Table 1. For the representation of the unknowns (μ_a, μ_s'), the domain Ω was divided into $n_p = 7{,}668$ cubic voxels (i.e., number of unknowns $n = 15{,}336$) in both models $A_{\Omega,\delta}$ and $A_{\Omega,h}$. Thus, the projector $P : x_\delta \mapsto x_h$ between the models is the identity matrix.

Construction of Error Models

To construct the enhanced error model, we proceeded as in section "Approximation Error Approach." The size of the random ensemble S from the prior model $\pi(x_\delta)$, Eq. 120, was $L = 384$. Figure 8 shows central xy-slices from two realizations of absorption and scatter images from the ensemble (the location of the slice is denoted by green line in Fig. 7). Using the ensemble, Gaussian approximations $\varepsilon \sim \mathcal{N}(\varepsilon_*, \Gamma_\varepsilon)$ for the error between the accurate model $A_{\Omega,\delta}$ and the target models were computed.

To assess the magnitude of the modeling error, we estimate signal-to-noise (SNR) ratio of the modeling error as

$$\mathrm{SNR} = 10 \log_{10} \left(\frac{\|\overline{A_{\Omega,\delta}}\|^2}{\|\varepsilon_*\|^2 + \mathrm{trace}(\Gamma_\varepsilon)} \right),$$

where $\overline{A_{\Omega,\delta}}$ is the mean of the accurate model $A_{\Omega,\delta}$ over the ensemble S. The SNR is estimated separately for the amplitude and phase part of the forward model.

Consider now the modeling error between the accurate model $A_{\Omega,\delta}$ ($N_n = 148{,}276$ nodes) and target model $A_{\Omega,h}$ ($N_n = 2{,}413$ nodes) in the first test case. In this case, the estimated SNRs for the modeling error in log amplitude and phase are approximately 20 and 13, corresponding to error levels of 10 and 22 %, respectively. These error levels exceed clearly typical levels of measurement noise in DOT measurements. Left image in Fig. 9 displays the covariance matrix Γ_ε, revealing the correlation structure of ε. Combining the high magnitude and complicated correlation structure of the modeling error ε with the fact that the inverse problem is sensitive to modeling errors, one can expect significant artifacts in the reconstructions with conventional noise model when employing the target model $A_{\Omega,h}$.

Right image in Fig. 9 shows normalized eigenvalues $\lambda/\lambda_{\max}$ of Γ_ε for the modeling error between models $A_{\Omega,\delta}$ and $A_{\Omega,h}$ in the first test case. As can be seen, the eigenvalues are decaying rapidly and already the 40th eigenvalue is less than 1 % of the maximum. Roughly speaking, this rapid decay of the eigenvalues can be interpreted such that the variability in the modeling error can be well explained with a relatively small number of principal components. In other words, one can take this as a sign that the structure of the modeling error is not "heavily dependent" on the realization of x or the prior model $\pi(x)$, and thus, the error model can be expected to perform well.

Notice that the setting up of the error model is a computationally intensive task, while the use of the model is as with the conventional error model. The computation time for setting up the error model is roughly equivalent to the size of the ensemble times, the time for forward solution in the accurate and approximate models. However, the error model needs to be estimated only once for a fixed measurement setup, and this estimation can be offline.

Computation of the MAP Estimates

The MAP-CEM and MAP-EEM estimates are computed by a Polak Ribiere conjugate gradient algorithm which is equipped with an explicit line search. Similarly as in the initial estimation, the positivity prior of the absorption and scatter images is taken into account by using (scaled) logarithmic parameterization

$$\log\left(\frac{\mu_\mathrm{a}}{\mu_{\mathrm{a}0}}\right), \quad \log\left(\frac{\mu'_\mathrm{s}}{\mu_{\mathrm{s}0}}\right)$$

in the unconstrained optimization process; for details, see [88].

Results are shown in Figs. 10 and 11 and computation times in Table 2.

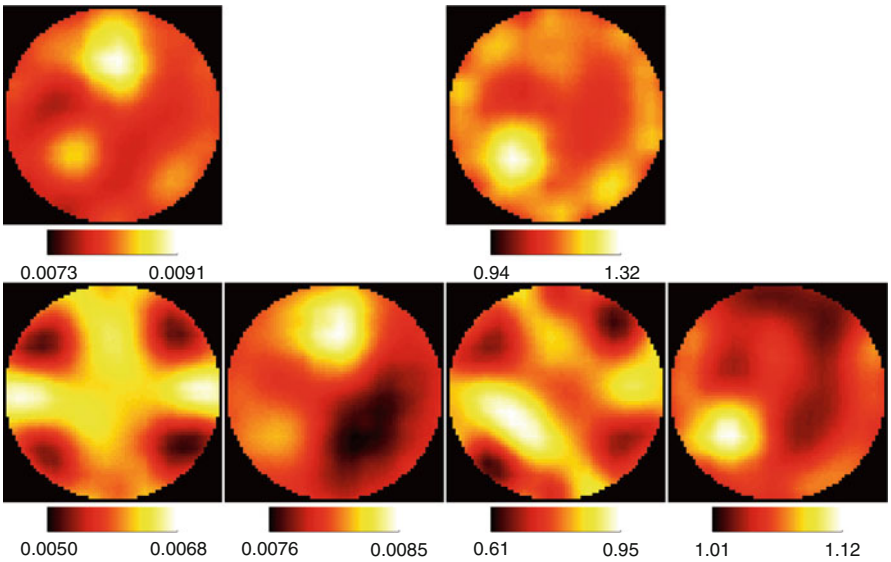


Fig. 10 Pure discretization errors. Central horizontal slice from the 3D reconstructions of absorption μ_a and scattering μ_s'. *Top row*: MAP estimate with the conventional error model (MAP-CEM) using the accurate forward model $A_{\Omega,\delta}$ (number of nodes in the FEM mesh $N_n = 148{,}276$). *Left*: $\mu_{a,CEM}$. *Right*: $\mu_{a,EEM}$. *Bottom row*: MAP estimates with the conventional (MAP-CEM) and enhanced error models (MAP-EEM) using the target model $A_{\Omega,h}$ (the number of nodes $N_n = 2{,}413$). Correct model domain $\Omega = \Omega$ is used in the target model $A_{\Omega,h}$. *Columns from left to right*: $\mu_{a,CEM}$, $\mu_{a,EEM}$, $\mu_{s,CEM}$, and $\mu_{s,EEM}$. The number of unknowns $x = (\mu_a, \mu_s')^T$ in the estimation with both models, $A_{\Omega,\delta}$ and $A_{\Omega,h}$, was 15,336

The images in the top row display the MAP estimate with the conventional noise model using the accurate forward model $A_{\Omega,\delta}$ ($N_n = 148{,}276$). The estimated values of global parameters $\mu_{a0} = 0.0079\,\text{mm}^{-1}$ and $\mu_{s0} = 1.086\,\text{mm}^{-1}$ are relatively close to the background values $\mu_{a,bg} = 0.01\,\text{mm}^{-1}$ and $\mu_{s,bg} = 1\,\text{mm}^{-1}$ of the target phantom. As can be seen, the structure of the phantom is reconstructed well, but the contrast of the recovered inclusions is low compared to the (presumed) contrast. However, the low contrast is related to the measurement setup, not the reconstruction algorithm; the same measurement realization has previously been used for absolute reconstructions with different algorithm in [88], resulting to similar reconstruction quality and contrast in the optical properties. See also [89] for similar results with the same measurement system. The MAP-CEM estimate with the accurate model $A_{\Omega,\delta}$ can be considered here as a *reference estimate* using conventional noise model in absence of modeling errors caused by reduced discretization or domain truncation.

The MAP-CEM estimate using the coarse target model $A_{\Omega,h}$ ($N_n = 2{,}413$) is shown in the first and third images in the bottom row in Figs. 10 and 11. As can be seen, the use of reduced discretization has caused significant errors in the reconstruction and also the levels of μ_a and μ_s' are erroneous.

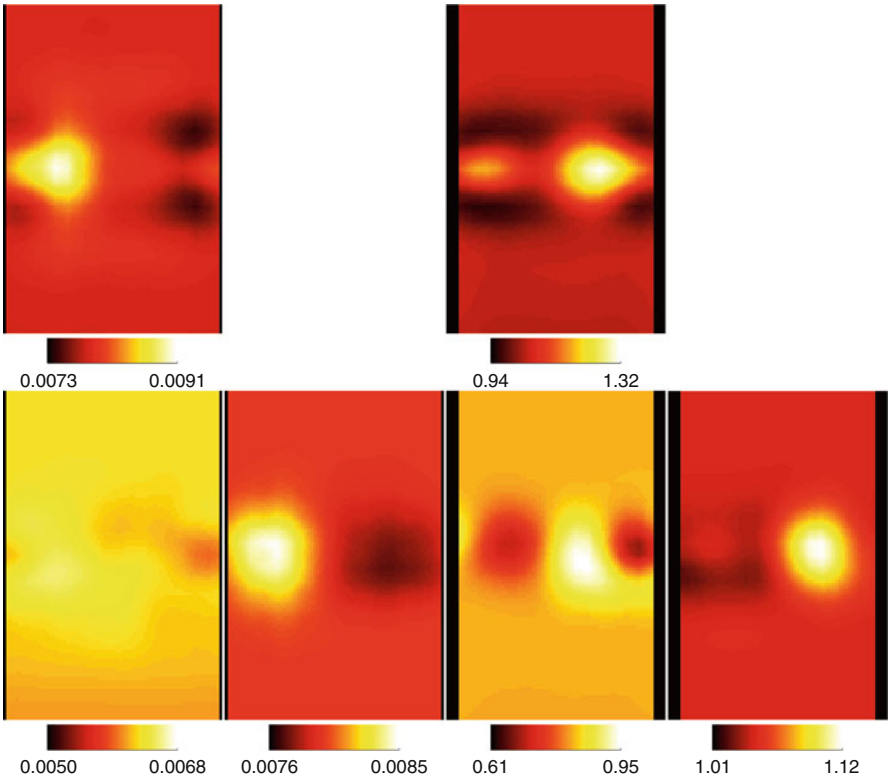


Fig. 11 Pure discretization errors. Vertical slices from the 3D reconstructions of absorption μ_a and scattering μ_s'. The slices have been chosen such that the inclusion in the parameter is visible. The arrangement of the images is equivalent to Fig. 10

Table 2 Reconstruction times for Figs. 10 and 11. t_{init} is the (wall clock) time for initial estimation, t_{MAP} for the MAP estimation, and t_{tot} the total reconstruction time (initial + MAP)

Noise model	Forward model	t_{init} (s)	t_{MAP} (s)	t_{tot} (s)
CEM	$A_{\Omega,\delta}$	126 min 20 s	173 min 22 s	299 min 44 s
CEM	$A_{\Omega,h}$	1 min 11 s	7 min 18 s	8 min 29 s
EEM	$A_{\Omega,h}$	28 s	7 min 34 s	8 min 2 s

The MAP estimate with the enhanced error model using the coarse target model $A_{\Omega,h}$ is shown in the second and fourth images in the bottom row in Figs. 10 and 11. As can be seen, the estimate is very similar to the MAP-CEM estimate with the accurate model $A_{\Omega,\delta}$, showing that the use of enhanced error model has efficiently compensated for the errors caused by reduced discretization accuracy. These results indicate that the enhanced error model allows significant reduction in computation time without compromise in the reconstruction quality; whereas the reconstruction time for the MAP-CEM using accurate model $A_{\Omega,\delta}$ is very close to 5 h, the computation time for MAP-EEM is only 8 min.

5 Conclusion

In this chapter, we mainly discussed the use of the diffusion approximation for optical tomography. Because of the exponentially ill-posed nature of the corresponding inverse problem, diffuse optical tomography (DOT) gives low resolution images. Current research is focused on several areas: the use of auxiliary (multimodality) information to improve DOT images, the development of smaller-scale (mesoscopic) imaging methods based on the radiative transfer equation, and the development of fluorescence and bioluminescence imaging techniques which give stronger contrast to features of interest. These methods are closely tied to the development of new experimental systems and to application areas which are driving the continued interest in this technique.

Cross-References

- ▸ Electrical Impedance Tomography
- ▸ Imaging in Random Media
- ▸ Inverse Scattering
- ▸ Photoacoustic and Thermoacoustic Tomography: Image Formation Principles
- ▸ Regularization Methods for Ill-Posed Problems

References

1. Ackroyd, R.T.: Finite Element Methods for Particle Transport: Applications to Reactor and Radiation Physics. Research Studies, Taunton (1997)
2. Amaldi, E.: The production and slowing down of neutrons. In: Flügge, S. (ed.) Encyclopedia of Physics, vol. 38/2, pp. 1–659. Springer, Berlin (1959)
3. Anderson, B.D.O., Moore, J.B.: Optimal Filtering. Prentice Hall, Englewood Cliffs (1979)
4. Aronson, R.: Boundary conditions for diffusion of light. J. Opt. Soc. Am. A **12**, 2532–2539 (1995)
5. Arridge, S.R.: Optical tomography in medical imaging. Inverse Probl. **15**(2), R41–R93 (1999)
6. Arridge, S.R., Lionheart, W.R.B.: Non-uniqueness in diffusion-based optical tomography. Opt. Lett. **23**, 882–884 (1998)
7. Arridge, S.R., Schotland, J.C.: Optical tomography: forward and inverse problems. Inverse Probl. **25**(12), 123010 (59pp) (2009)
8. Arridge, S.R., Cope, M., Delpy, D.T.: Theoretical basis for the determination of optical pathlengths in tissue: temporal and frequency analysis. Phys. Med. Biol. **37**, 1531–1560 (1992)
9. Arridge, S.R., Schweiger, M., Hiraoka, M., Delpy, D.T.: A finite element approach for modeling photon transport in tissue. Med. Phys. **20**(2), 299–309 (1993)
10. Arridge, S.R., Dehghani, H., Schweiger, M., Okada, E.: The finite element model for the propagation of light in scattering media: a direct method for domains with non-scattering regions. Med. Phys. **27**(1), 252–264 (2000)
11. Arridge, S.R., Kaipio, J.P., Kolehmainen, V., Schweiger, M., Somersalo, E., Tarvainen, T., Vauhkonen, M.: Approximation errors and model reduction with an application in optical diffusion tomography. Inverse Probl. **22**(1), 175–196 (2006)

12. Arridge, S.R., Kaipio, J.P., Kolehmainen, V., Schweiger, M., Somersalo, E., Tarvainen, T., Vauhkonen, M.: Approximation errors and model reduction with an application in optical diffusion tomography. Inverse Probl. **22**, 175–195 (2006)
13. Aydin, E.D.: Three-dimensional photon migration through voidlike regions and channels. Appl. Opt. **46**(34), 8272–8277 (2007)
14. Aydin, E.D., de Oliveira, C.R.E., Goddard, A.J.H.: A finite element-spherical harmonics radiation transport model for photon migration in turbid media. J. Quant. Spectrosc. Radiat. Transf. **84**, 247–260 (2004)
15. Bal, G.: Transport through diffusive and nondiffusive regions, embedded objects, and clear layers. SIAM J. Appl. Math. **62**(5), 1677–1697 (2002)
16. Bal, G.: Radiative transfer equation with varying refractive index: a mathematical perspective. J. Opt. Soc. Am. A **23**, 1639–1644 (2006)
17. Bal, G.: Inverse transport theory and applications. Inverse Probl. **25**, 053001 (48pp) (2009)
18. Bal, G., Maday, Y.: Coupling of transport and diffusion models in linear transport theory. Math. Model. Numer. Anal. **36**(1), 69–86 (2002)
19. Benaron, D.A., Stevenson, D.K.: Optical time-of-flight and absorbance imaging of biological media. Science **259**, 1463–1466 (1993)
20. Berg, R., Svanberg, S., Jarlman, O.: Medical transillumination imaging using short-pulse laser diodes. Appl. Opt. **32**, 574–579 (1993)
21. Berger, J.O.: Statistical Decision Theory and Bayesian Analysis. Springer, New York (2006)
22. Bluestone, A.V., Abdoulaev, G., Schmitz, C.H., Barbour, R.L., Hielscher, A.H.: Three-dimensional optical tomography of hemodynamics in the human head. Opt. Express **9**(6), 272–286 (2001)
23. Calvetti, D., Kaipio, J.P., Somersalo, E.: Aristotelian prior boundary conditions. Int. J. Math. **1**, 63–81 (2006)
24. Case, M.C., Zweifel, P.F.: Linear Transport Theory. Addison-Wesley, New York (1967)
25. Contini, D., Martelli, F., Zaccanti, G.: Photon migration through a turbid slab described by a model based on diffusion approximation. I. Theory Appl. Opt. **36**(19), 4587–4599 (1997)
26. Cope, M., Delpy, D.T.: System for long term measurement of cerebral blood and tissue oxygenation on newborn infants by near infra-red transillumination. Med. Biol. Eng. Comput. **26**, 289–294 (1988)
27. Cutler, M.: Transillumination as an aid in the diagnosis of breast lesions. Surg. Gynecol. Obstet. **48**, 721–729 (1929)
28. Dehghani, H., Arridge, S.R., Schweiger, M., Delpy, D.T.: Optical tomography in the presence of void regions. J. Opt. Soc. Am. A **17**(9), 1659–1670 (2000)
29. Delpy, D.T., Cope, M., van der Zee, P., Arridge, S.R., Wray, S., Wyatt, J.: Estimation of optical pathlength through tissue from direct time of flight measurement. Phys. Med. Biol. **33**, 1433–1442 (1988)
30. Diamond, S.G., Huppert, T.J., Kolehmainen, V., Franceschini, M.A., Kaipio, J.P., Arridge, S.R., Boas, D.A.: Dynamic physiological modeling for functional diffuse optical tomography. Neuroimage **30**, 88–101 (2006)
31. Dorn, O.: Das inverse Transportproblem in der Lasertomographie. PhD thesis, University of Münster (1997)
32. Doucet, A., de Freitas, N., Gordon, N.: Sequential Monte Carlo Methods in Practice. Springer, New York (2001)
33. Duderstadt, J.J., Martin, W.R.: Transport Theory. Wiley, New York (1979)
34. Durbin, J., Koopman, J.: Time Series Analysis by State Space Methods. Oxford University Press, Oxford (2001)
35. Fantini, S., Franceschini, M.A., Gratton, E.: Effective source term in the diffusion equation for photon transport in turbid media. Appl. Opt. **36**(1), 156–163 (1997)
36. Ferwerda, H.A.: The radiative transfer equation for scattering media with a spatially varying refractive index. J. Opt. A Pure Appl. Opt. **1**(3), L1–L2 (1999)

37. Firbank, M., Arridge, S.R., Schweiger, M., Delpy, D.T.: An investigation of light transport through scattering bodies with non-scattering regions. Phys. Med. Biol. **41**, 767–783 (1996)
38. Furutsu, K.: Diffusion equation derived from space-time transport equation. J. Opt. Soc. Am. **70**(4), 360–366 (1980)
39. Groenhuis, R.A.J., Ferwerda, H.A., Ten Bosch, J.J.: Scattering and absorption of turbid materials determined from reflection measurements. Part 1: theory. Appl. Opt. **22**(16), 2456–2462 (1983)
40. Haskell, R.C., Svaasand, L.O., Tsay, T.-T., Feng, T.-C., McAdams, M.S., Tromberg, B.J.: Boundary conditions for the diffusion equation in radiative transfer. J. Opt. Soc. Am. A **11**(10), 2727–2741 (1994)
41. Hayashi, T., Kashio, Y., Okada, E.: Hybrid Monte Carlo-diffusion method for light propagation in tissue with a low-scattering region. Appl. Opt. **42**(16), 2888–2896 (2003)
42. Hebden, J.C., Kruger, R.A., Wong, K.S.: Time resolved imaging through a highly scattering medium. Appl. Opt. **30**(7), 788–794 (1991)
43. Hebden, J.C., Gibson, A., Md Yusof, R., Everdell, N., Hillman, E.M.C., Delpy, D.T., Arridge, S.R., Austin, T., Meek, J.H., Wyatt, J.S.: Three-dimensional optical tomography of the premature infant brain. Phys. Med. Biol. **47**, 4155–4166 (2002)
44. Heino, J., Somersalo, E.: Estimation of optical absorption in anisotropic background. Inverse Probl. **18**, 559–573 (2002)
45. Heino, J., Somersalo, E.: A modelling error approach for the estimation of optical absorption in the presence of anisotropies. Phys. Med. Biol. **49**, 4785–4798 (2004)
46. Heino, J., Somersalo, E., Kaipio, J.P.: Compensation for geometric mismodelling by anisotropies in optical tomography. Opt. Express **13**(1), 296–308 (2005)
47. Henyey, L.G., Greenstein, J.L.: Diffuse radiation in the galaxy. AstroPhys. J. **93**, 70–83 (1941)
48. Hielscher, A.H., Alcouffe, R.E., Barbour, R.L.: Comparison of finitedifference transport and diffusion calculations for photon migration in homogeneous and hetergeneous tissue. Phys. Med. Biol. **43**, 1285–1302 (1998)
49. Ho, P.P., Baldeck, P., Wong, K.S., Yoo, K.M., Lee, D., Alfano, R.R.: Time dynamics of photon migration in semiopaque random media. Appl. Opt. **28**, 2304–2310 (1989)
50. Huttunen, J.M.J., Kaipio, J.P.: Approximation error analysis in nonlinear state estimation with an application to state-space identification. Inverse Probl. **23**, 2141–2157 (2007)
51. Huttunen, J.M.J., Kaipio, J.P.: Approximation errors in nostationary inverse problems. Inverse Probl. Imaging **1**(1), 77–93 (2007)
52. Huttunen, J.M.J., Kaipio, J.P.: Model reduction in state identification problems with an application to determination of thermal parameters. Appl. Numer. Math. **59**, 877–890 (2009)
53. Huttunen, J.M.J., Lehikoinen, A., Hämäläinen, J., Kaipio, J.P.: Importance filtering approach for the nonstationary approximation error method. Inverse Probl. (2009). In review
54. Ishimaru, A.: Wave Propagation and Scattering in Random Media, vol. 1. Academic, New York (1978)
55. Jarry, G., Ghesquiere, S., Maarek, J.M., Debray, S., Bui, M.-H., Laurent, H.D.: Imaging mammalian tissues and organs using laser collimated transillumination. J. Biomed. Eng. **6**, 70–74 (1984)
56. Jöbsis, F.F.: Noninvasive infrared monitoring of cerebral and myocardial oxygen sufficiency and circulatory parameters. Science **198**, 1264–1267 (1977)
57. Kaipio, J., Somersalo, E.: Statistical and Computational Inverse Problems. Springer, New York (2005)
58. Kaipio, J., Somersalo, E.: Statistical and computational inverse problems. J. Comput. Appl. Math. **198**, 493–504 (2007)
59. Kaipio, J.P., Kolehmainen, V., Vauhkonen, M., Somersalo, E.: Inverse problems with structural prior information. Inverse Probl. **15**, 713–729 (1999)
60. Kak, A.C., Slaney, M.: Principles of Computerized Tomographic Imaging. IEEE, New York (1987)

61. Kalman, R.E.: A new approach to linear filtering and prediction problems. Trans ASME J. Basic Eng. **82D**(1), 35–45 (1960)
62. Khan, T., Jiang, H.: A new diffusion approximation to the radiative transfer equation for scattering media with spatially varying refractive indices. J. Opt. A Pure Appl. Opt. **5**, 137–141 (2003)
63. Kim, A.D., Ishimaru, A.: Optical diffusion of continuos-wave, pulsed, and density waves in scattering media and comparisons with radiative transfer. Appl. Opt. **37**(22), 5313–5319 (1998)
64. Klose, A.D., Larsen, E.W.: Light transport in biological tissue based on the simplified spherical harmonics equations. J. Comput. Phys. **220**, 441–470 (2006)
65. Kolehmainen, V., Arridge, S.R., Vauhkonen, M., Kaipio, J.P.: Simultaneous reconstruction of internal tissue region boundaries and coefficients in optical diffusion tomography. Phys. Med. Biol. **45**, 3267–3283 (2000)
66. Kolehmainen, V., Prince, S., Arridge, S.R., Kaipio, J.P.: A state estimation approach to non-stationary optical tomography problem. J. Opt. Soc. Am. A **20**, 876–884 (2000)
67. Kolehmainen, V., Schweoger, M., Nissilä, I., Tarvainen, T., Arridge, S.R., Kaipio, J.P.: Approximation errors and model reduction in three-dimensional optical tomography. J. Opt. Soc. Am. A **26**, 2257–2268 (2009)
68. Kolehmainen, V., Tarvainen, T., Arridge, S.R., Kaipio, J.P.: Marginalization of uninteresting distributed parameters in inverse problems – application to diffuse optical tomography. Int. J. Uncertain. Quantif. (2010, in press)
69. Lakowicz, J.R., Berndt, K.: Frequency domain measurement of photon migration in tissues. Chem. Phys. Lett. **166**(3), 246–252 (1990)
70. Lehikoinen, A., Finsterle, S., Voutilainen, A., Heikkinen, L.M., Vauhkonen, M., Kaipio, J.P.: Approximation errors and truncation of computational domains with application to geophysical tomography. Inverse Probl. Imaging **1**, 371–389 (2007)
71. Lehikoinen, A., Huttunen, J.M.J., Finsterle, S., Kowalsky, M.B., Kaipio, J.P.: Dynamic inversion for hydrological process monitoring with electrical resistance tomography under model uncertainties. Water Resour. Res. **46**, W04513 (2010). doi:10.1029/2009WR008470
72. Marti-Lopez, L., Bouza-Dominguez, J., Hebden, J.C., Arridge, S.R., Martinez-Celorio, R.A.: Validity conditions for the radiative transfer equation. J. Opt. Soc. Am. A **20**(11), 2046–2056 (2003)
73. Mitic, G., Kolzer, J., Otto, J., Plies, E., Solkner, G., Zinth, W.: Timegated transillumination of biological tissue and tissuelike phantoms. Opt. Lett. **33**, 6699–6710 (1994)
74. Natterer, F., Wübbeling, F.: Mathematical Methods in Image Reconstruction. SIAM, Philadelphia (2001)
75. Nissilä, I., Noponen, T., Kotilahti, K., Tarvainen, T., Schweiger, M., Lipiänen, L., Arridge, S.R., Katila, T.: Instrumentation and calibration methods for the multichannel measurement of phase and amplitude in optical tomography. Rev. Sci. Instrum. **76**(4), 004302 (2005)
76. Nissinen, A., Heikkinen, L.M., Kaipio, J.P.: Approximation errors in electrical impedance tomography – an experimental study. Meas. Sci. Technol. **19** (2008). doi:10.1088/0957-0233/19/1/015501
77. Nissinen, A., Heikkinen, L.M., Kolehmainen, V., Kaipio, J.P.: Compensation of errors due to discretization, domain truncation and unknown contact impedances in electrical impedance tomography. Meas. Sci. Technol. **20** (2009). doi:10.1088/0957–0233/20/10/105504
78. Nissinen, A., Kolehmainen, V., Kaipio, J.P.: Compensation of modelling errors due to unknown domain boundary in electrical impedance tomography. IEEE Trans. Med. Imaging (2010). In review
79. Ntziachristos, V., Ma, X., Chance, B.: Time-correlated single photon counting imager for simultaneous magnetic resonance and near-infrared mammography. Rev. Sci. Instrum. **69**, 4221–4233 (1998)
80. Okada, E., Schweiger, M., Arridge, S.R., Firbank, M., Delpy, D.T.: Experimental validation of Monte Carlo and Finite-Element methods for the estimation of the optical path length in inhomogeneous tissue. Appl. Opt. **35**(19), 3362–3371 (1996)

81. Prince, S., Kolehmainen, V., Kaipio, J.P., Franceschini, M.A., Boas, D., Arridge, S.R.: Time series estimation of biological factors in optical diffusion tomography. Phys. Med. Biol. **48**(11), 1491–1504 (2003)
82. Schmidt, A., Corey, R., Saulnier, P.: Imaging through random media by use of low-coherence optical heterodyning. Opt. Lett. **20**, 404–406 (1995)
83. Schmidt, F.E.W., Fry, M.E., Hillman, E.M.C., Hebden, J.C., Delpy, D.T.: A 32-channel time-resolved instrument for medical optical tomography. Rev. Sci. Instrum. **71**(1), 256–265 (2000)
84. Schmitt, J.M., Gandbjbakhche, A.H., Bonner, R.F.: Use of polarized light to discriminate short-path photons in a multiply scattering medium. Appl. Opt. **31**, 6535–6546 (1992)
85. Schotland, J.C., Markel, V.: Inverse scattering with diffusing waves. J. Opt. Soc. Am. A **18**, 2767–2777 (2001)
86. Schweiger, M., Arridge, S.R.: The finite element model for the propagation of light in scattering media: frequency domain case. Med. Phys. **24**(6), 895–902 (1997)
87. Schweiger, M., Arridge, S.R., Hiraoka, M., Delpy, D.T.: The finite element model for the propagation of light in scattering media: boundary and source conditions. Med. Phys. **22**(11), 1779–1792 (1995)
88. Schweiger, M., Arridge, S.R., Nissilä, I.: Gauss–Newton method for image reconstruction in diffuse optical tomography. Phys. Med. Biol. **50**, 2365–2386 (2005)
89. Schweiger, M., Nissilä, I., Boas, D.A., Arridge, S.R.: Image reconstruction in optical tomography in the presence of coupling errors. Appl. Opt. **46**(14), 2743–2756 (2007)
90. Spears, K.G., Serafin, J., Abramson, N.H., Zhu, X., Bjelkhagen, H.: Chronocoherent imaging for medicine. IEEE Trans. Biomed. Eng. **36**, 1210–1221 (1989)
91. Sylvester, J., Uhlmann, G.: A global uniquness theorem for an inverse boundary value problem. Ann. Math. **125**, 153–169 (1987)
92. Tarvainen, T., Vauhkonen, M., Kolehmainen, V., Arridge, S.R., Kaipio, J.P.: Coupled radiative transfer equation and diffusion approximation model for photon migration in turbid medium with low-scattering and non-scattering regions. Phys. Med. Biol. **50**, 4913–4930 (2005)
93. Tarvainen, T., Vauhkonen, M., Kolehmainen, V., Kaipio, J.P.: A hybrid radiative transfer – diffusion model for optical tomography. Appl. Opt. **44**(6), 876–886 (2005)
94. Tarvainen, T., Vauhkonen, M., Kolehmainen, V., Kaipio, J.P.: Finite element model for the coupled radiative transfer equation and diffusion approximation. Int. J. Numer. Methods Eng. **65**(3), 383–405 (2006)
95. Tarvainen, T., Kolehmainen, V., Pulkkinen, A., Vauhkonen, M., Schweiger, M., Arridge, S.R., Kaipio, J.P.: Approximation error approach for compensating for modelling errors between the radiative transfer equation and the diffusion approximation in diffuse optical tomography. Inverse Probl. **26** (2010). doi:10.1088/0266–5611/26/1/015005
96. Tervo, J., Kolmonen, P., Vauhkonen, M., Heikkinen, L.M., Kaipio, J.P.: A finite-element model of electron transport in radiation therapy and a related inverse problem. Inverse Probl. **15**, 1345–1362 (1999)
97. Wang, L.V.: Rapid modeling of diffuse reflectance of light in turbid slabs. J. Opt. Soc. Am. A **15**(4), 936–944 (1998)
98. Wang, L., Jacques, S.L.: Hybrid model of Monte Carlo simulation diffusion theory for light reflectance by turbid media. J. Opt. Soc. Am. A **10**(8), 1746–1752 (1993)
99. Wang, L., Ho, P.P., Liu, C., Zhang, G., Alfano, R.R.: Ballistic 2-D imaging through scattering walls using an ultrafast optical Kerr gate. Science **253**, 769–771 (1991)
100. Wright, S., Schweiger, M., Arridge, S.R.: Reconstruction in optical tomography using the PN approximations. Meas. Sci. Technol. **18**, 79–86 (2007)

Photoacoustic and Thermoacoustic Tomography: Image Formation Principles

Kun Wang and Mark A. Anastasio

Contents

1 Introduction 1082
2 Imaging Physics and Contrast Mechanisms 1083
The Thermoacoustic Effect and Signal Generation 1083
Image Contrast in Laser-Based PAT 1086
Image Contrast in RF-Based PAT 1087
Functional PAT 1088
3 Principles of PAT Image Reconstruction 1089
PAT Imaging Models in Their Continuous Forms 1091
Universal Backprojection Algorithm 1092
The Fourier-Shell Identity 1093
Spatial Resolution from a Fourier Perspective 1095
4 Speed-of-Sound Heterogeneities and Acoustic Attenuation 1099
Frequency-Dependent Acoustic Attenuation 1099
Weak Variations in the Speed-of-Sound Distribution 1100
5 Data Redundancies and the Half-Time Reconstruction Problem 1102
Data Redundancies 1102
Mitigation of Image Artifacts Due to Acoustic Heterogeneities 1103
6 Discrete Imaging Models 1104
Continuous-to-Discrete Imaging Models 1105
Finite-Dimensional Object Representations 1107
Discrete-to-Discrete Imaging Models 1108
Iterative Image Reconstruction 1110
7 Conclusion 1113
Cross-References 1113
References 1113

K. Wang (✉)
Medical Imaging Research Center, Illinois Institute of Technology, Chicago, IL, USA
e-mail: kwang 16@iit.edu

M.A. Anastasio
Biomedical Engineering, Illinois Institute of Technology, Chicago, IL, USA
e-mail: anastasio@iit.edu

© Springer Science+Business Media New York 2015

O. Scherzer (ed.), *Handbook of Mathematical Methods in Imaging*,
DOI 10.1007/978-1-4939-0790-8_50

Abstract

Photoacoustic tomography (PAT), also known as thermoacoustic or optoacoustic tomography, is a rapidly emerging imaging technique that holds great promise for biomedical imaging. PAT is a hybrid imaging technique, and can be viewed either as an ultrasound mediated electromagnetic modality or an ultrasound modality that exploits electromagnetic-enhanced image contrast. In this chapter, we provide a review of the underlying imaging physics and contrast mechanisms in PAT. Additionally, the imaging models that relate the measured photoacoustic wavefields to the sought-after optical absorption distribution are described in their continuous and discrete forms. The basic principles of image reconstruction from discrete measurement data are presented, which includes a review of methods for modeling the measurement system response.

1 Introduction

Photoacoustic tomography (PAT), also known as thermoacoustic or optoacoustic tomography, is a rapidly emerging imaging technique that holds great promise for biomedical imaging [30, 31, 45, 61, 63]. PAT is a hybrid technique that exploits the thermoacoustic effect for signal generation. It seeks to combine the high electromagnetic contrast of tissue with the high spatial resolution of ultrasonic methods. Accordingly, PAT can be viewed either as an ultrasound mediated electromagnetic modality or an ultrasound modality that exploits electromagnetic-enhanced image contrast [65]. Since the 1990s, there have been numerous fundamental studies of photoacoustic imaging of biological tissue [19, 31, 41, 44, 46, 47, 57], and the development of PAT continues to progress at a tremendous rate [18, 24, 28, 30, 31, 33, 45, 64, 65].

When a short electromagnetic pulse (e.g., microwave or laser) is used to irradiate a biological tissue, the thermoacoustic effect results in the emission of acoustic signals that can be measured outside the object by use of wideband ultrasonic transducers. The objective of PAT is to produce an image that represents a map of the spatially variant electromagnetic absorption properties of the tissue, from knowledge of the measured acoustic signals. Because the optical absorption properties of tissue are highly related to its molecular constitution, PAT images can reveal the pathological condition of the tissue [11, 26] and therefore facilitate a wide range of diagnostic tasks. Moreover, when employed with targeted probes or optical contrast agents, PAT has the potential to facilitate high-resolution molecular imaging [32, 58] of deep structures, which cannot be achieved easily with pure optical methods.

From a physical perspective, the image reconstruction problem in PAT can be interpreted as an inverse source problem [6]. Accordingly, PAT is a computed imaging modality that utilizes an image reconstruction algorithm to form the image of the absorbed optical energy distribution. A variety of analytic image reconstruction algorithms have been developed for three-dimensional (3D) PAT,

assuming point-like ultrasound transducers with canonical measurement apertures [22, 23, 30, 31, 35, 64–66]. All known analytic reconstruction algorithms that are mathematically exact and numerically stable require complete knowledge of the photoacoustic wavefield on a measurement aperture that either encloses the entire object or extends to infinity. In many potential applications of PAT imaging, it is not feasible to acquire such measurement data. Because of this, iterative, or more generally, optimization-based, reconstruction algorithms for PAT are being developed actively [4, 5, 18, 48, 50] that provide the opportunity for accurate image reconstruction from incomplete measurement data. Iterative reconstruction algorithms also allow for accurate modeling of physical nonidealities in the data, such as those introduced by acoustic inhomogeneity and attenuation, or the response of the imaging system.

In this chapter, the physical principles of PAT are reviewed. We start with a review of the underlying imaging physics and contrast mechanisms in PAT. Subsequently, the imaging models that relate the measured photoacoustic wavefields to the sought-after optical absorption distribution are described in their continuous and discrete forms. The basic principles of image reconstruction from discrete measurement data are presented, which includes a review of methods for modeling the measurement system response. We defer a detailed description of analytic reconstruction algorithms and the mathematical properties of PAT in ▸ Mathematics of Photoacoustic and Thermoacoustic Tomography.

2 Imaging Physics and Contrast Mechanisms

In PAT, a laser or microwave source is used to irradiate an object, and the thermoacoustic effect results in the generation of a pressure wavefield $p(\mathbf{r}, t)$ [45, 54, 65], where $\mathbf{r} \in \mathbb{R}^3$ and t is the temporal coordinate. The resulting pressure wavefield can be measured by use of wideband ultrasonic transducers located on a measurement aperture $\Omega_0 \subset \mathbb{R}^3$, which is a 2D surface that partially or completely surrounds the object. In this section, we review the physics that underlies the image contrast mechanism in PAT employing laser and microwave sources.

The Thermoacoustic Effect and Signal Generation

The generation of photoacoustic wavefields in an inviscid and lossless medium is described by the general photoacoustic wave equation [55, 56]

$$\left(\nabla^2 - \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\right) p(\mathbf{r}, t) = -\frac{\beta}{\kappa c^2}\frac{\partial^2 T(\mathbf{r}, t)}{\partial t^2}, \tag{1}$$

where ∇^2 is the 3D Laplacian operator, $T(\mathbf{r}, t)$ denotes the temperature rise within the object at location $\mathbf{r}$ and time t due to absorption of the probing electromagnetic radiation, and $p(\mathbf{r}, t)$ denotes the resulting induced acoustic pressure. The quantities

β, κ, and c denote the thermal coefficient of volume expansion, isothermal compressibility, and speed of sound, respectively. Because an inviscid medium is assumed, the propagation of shear waves is neglected in Eq. (1), which is typically reasonable for soft-tissue imaging applications. Note that the spatial–temporal samples of $p(\mathbf{r}, t)$, which are subsequently degraded by the response of the imaging system, represent the measurement data in a PAT experiment.

When the temporal width of the exciting electromagnetic pulse is sufficiently short, the pressure wavefield is produced before significant heat conduction can take place. In this situation, the excitation is said to be in thermal confinement. Specifically, this occurs when the temporal width τ of the exciting electromagnetic pulse satisfies [56]

$$\tau < \frac{d_c^2}{4\alpha_{th}}, \tag{2}$$

where d_c and α_{th} denote the characteristic dimension (m) of the heated region and the thermal diffusivity (m^2/s).

Under conditions of thermal confinement, the temperature function $T(\mathbf{r}, t)$ satisfies

$$\rho C_V \frac{\partial T(\mathbf{r}, t)}{\partial t} = H(\mathbf{r}, t), \tag{3}$$

where ρ and C_V denote the mass density (kg/m^3) and specific heat capacity of the medium at constant volume. The quantity $H(\mathbf{r}, t)$ [$J/(m^3 s)$] is called the heating function that describes the energy per unit volume and time that is deposited in the medium by the exciting electromagnetic pulse. On substitution from Eq. (3) into Eq. (1), one obtains the simplified photoacoustic wave equation

$$\left(\nabla^2 - \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\right) p(\mathbf{r}, t) = -\frac{\beta}{C_p}\frac{\partial H(\mathbf{r}, t)}{\partial t}, \tag{4}$$

where $C_p = \rho c^2 \kappa C_V$ [J/(kg K)] denotes the specific heat capacity of the medium at constant pressure. It is sometimes convenient to work the velocity potential $\phi(\mathbf{r}, t)$ that is related to the pressure as $p(\mathbf{r}, t) = -\rho \frac{\partial \phi(\mathbf{r}, t)}{\partial t}$. It can be readily verified that Eq. (4) can be reexpressed in terms of $\phi(\mathbf{r}, t)$ as

$$\left(\nabla^2 - \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\right) \phi(\mathbf{r}, t) = \frac{\beta}{\rho C_p} H(\mathbf{r}, t). \tag{5}$$

The photoacoustic wave equations described by Eqs. (4) and (5) have been solved for a variety of canonical absorbers [15–17]. Figure 1 shows an example corresponding to a uniform spherical absorber. In this case, the optical absorber was assumed to possess a speed of sound c_0 that matched the background medium. Note that the pressure possesses an "N-shape" waveform. Solutions have also been derived for the case where the optical absorbers have acoustical properties that are different from those of the background medium [15].

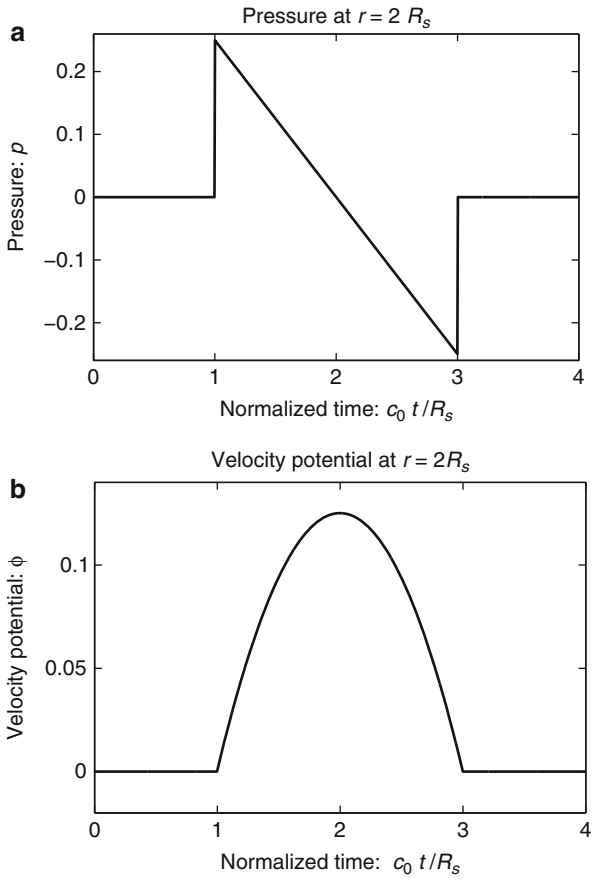


Fig. 1 The pressure (**a**) and velocity potential (**b**) waveforms produced by the thermoacoustic effect for a uniform sphere of radius R_s

In practice, it is appropriate to consider the following separable form for the heating function

$$H(\mathbf{r}, t) = A(\mathbf{r})I(t), \tag{6}$$

where $A(\mathbf{r})$ (J/m^3) is the absorbed energy density and $I(t)$ denotes the temporal profile of the illuminating pulse.

When the exciting electromagnetic pulse duration τ is short enough to satisfy the acoustic stress-confinement condition

$$\tau < \frac{d_c}{c}, \tag{7}$$

in addition to the thermal-confinement condition in Eq. (2), one can approximate $I(t)$ by a Dirac delta function $I(t) \approx \delta(t)$. Physically, Eq. (7) requires that all of the thermal energy has been deposited by the electromagnetic pulse before the mass

density or volume of the medium has had time to change. In this case, the absorbed energy density $A(\mathbf{r})$ is related to the induced pressure wavefield $p(\mathbf{r}, t)$ at $t = 0$ as

$$p(\mathbf{r}, t = 0) = \Gamma A(\mathbf{r}), \tag{8}$$

where Γ is the dimensionless Gruneisen parameter. As discussed in detail later, the goal of PAT is to determine $A(\mathbf{r})$ or, equivalently, $p(\mathbf{r}, t = 0)$ from measurements of $p(\mathbf{r}, t)$ acquired on a measurement aperture. It is also useful to note that under the acoustic stress-confinement condition, Eq. (4) coupled with appropriate boundary conditions is mathematically equivalent to the initial value problem [34]

$$\left(\nabla^2 - \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\right) p(\mathbf{r}, t) = 0, \tag{9}$$

subject to

$$p(\mathbf{r}, t = 0) = \Gamma A(\mathbf{r}) \quad \text{and} \quad \left.\frac{\partial p(\mathbf{r}, t)}{\partial t}\right|_{t=0} = 0. \tag{10}$$

The effects of heterogeneous speed of sound or acoustic attenuation are not addressed above, but will be described later. In the following two subsections, a review of the physical object properties that give rise to image contrast, that is, variations in $A(\mathbf{r})$, are reviewed for the case of optical and microwave illumination.

Image Contrast in Laser-Based PAT

When an optical laser pulse is employed to induce the thermoacoustic effect, the heating function can be explicitly expressed as

$$H(\mathbf{r}, t) = \mu_a(\mathbf{r})\Phi(\mathbf{r}, t), \tag{11}$$

where $\mu_a(\mathbf{r})$ (1/m) is the optical absorption coefficient of the medium and $\Phi(\mathbf{r}, t)$ [J/(m^{2}s)] is the optical fluence rate [39]. Assuming $\Phi(\mathbf{r}, t) \equiv \Phi_s(\mathbf{r})I(t)$, Eq. (11) can be expressed as

$$H(\mathbf{r}, t) = \underbrace{\mu_a(\mathbf{r})\Phi_s(\mathbf{r})}_{A(\mathbf{r})} I(t), \tag{12}$$

where the absorbed energy density, which is the sought-after quantity in PAT, is now identified as

$$A(\mathbf{r}) \equiv \mu_a(\mathbf{r})\Phi_s(\mathbf{r}). \tag{13}$$

Equation (13) reveals that image contrast in laser-based PAT is determined by the optical absorption properties of the object as well as variations in the fluence of

the illuminating optical radiation. Because only the optical absorption properties are intrinsic to the object, in implementation of PAT it is desirable to make the optical fluence $\Phi_s(\mathbf{r})$ as uniform as possible so one can unambiguously interpret $A(\mathbf{r}) \propto \mu_a(\mathbf{r})$. This presents experimental challenges, and computational methods for quantitative determination of $\mu_a(\mathbf{r})$ are being developed actively [9, 12, 68]. However, in most current implementations of PAT, an estimate of $A(\mathbf{r})$ represents the final image.

There are many desirable characteristics of laser-based PAT for biological imaging. The optical absorption coefficient $\mu_a(\mathbf{r})$ is a function of the molecular composition of tissue [11] and is therefore sensitive to tissue pathologies and functions. Specifically, PAT can deduce physiological parameters such as the oxygen saturation of hemoglobin and the total concentration of hemoglobin, as well as certain features of cancer such as elevated blood content of tissue due to angiogenesis [65].

Although pure optical imaging methods also are sensitive to such physiological parameters, they are limited by their relatively poor spatial resolution and inability to image deep tissue structures. PAT circumvents these limitations because diffusely scattered photons that are absorbed at deep locations are still useful for signal generation via the thermoacoustic effect. When the wavelength of the optical source lies in the range 700–900 nm, light can penetrate up to several centimeters in biological tissue. As described by Eq. (13), the optical fluence $\Phi_s(\mathbf{r})$, which contains ballistic and diffusely scattered photons, modulates $\mu_a(\mathbf{r})$. However, as described later, the spatial resolution of the reconstructed estimate of $A(\mathbf{r})$ is not directly affected by this and is determined largely by the properties of the measured pressure signal $p(\mathbf{r}, t)$.

Image Contrast in RF-Based PAT

When an RF pulse is employed to induce the thermoacoustic effect, the nature of the image contrast is different from that described above. A detailed analysis of this has been conducted by Li et al., in [40]. Consider the case of an RF pulse whose temporal width is much longer than the oscillation period of the electromagnetic wave at the center frequency ω_c. The RF source is assumed to produce a plane-wave with linear polarization and can be described as

$$e_{\text{in}}(t) = S(t)\cos(\omega_c t), \tag{14}$$

where $S(t)$ is a slowly varying envelope function. Furthermore, consider that the medium is isotropic and the electrical conductivity of the medium $\sigma(\mathbf{r}, \omega)$ can be approximated as

$$\sigma(\mathbf{r}, \omega) \approx \sigma(\mathbf{r}, \omega_c), \tag{15}$$

where ω represents the temporal frequency variable. Under the stated conditions, it is the short-time averaged heating function

$$< H(\mathbf{r},t) >\equiv \frac{1}{T_c}\int_t^{t+T_c} dt|H(\mathbf{r},t)|, \tag{16}$$

where $T_c = \frac{2\pi}{\omega_c}$, which gives rise to signal generation in RF-based PAT [40]. It has been demonstrated [40] that this quantity can be expressed as

$$< H(\mathbf{r},t) >= A(\mathbf{r})\frac{S^2(t)}{2}, \tag{17}$$

where $\frac{S^2(t)}{2}$ represents the electric field intensity of the RF source and

$$A(\mathbf{r}) \equiv \frac{\sigma(\mathbf{r},\omega_c)|\tilde{E}(\mathbf{r},\omega_c)|^2}{|\tilde{e}_{\text{in}}(\omega_c)|^2}, \tag{18}$$

where $\tilde{E}(\mathbf{r},\omega_c)$ and $\tilde{e}_{\text{in}}(\omega_c)$ denote the temporal Fourier transforms of $E(\mathbf{r},t)$ and $e_{\text{in}}(t)$ evaluated at $\omega = \omega_c$, with $E(\mathbf{r},t)$ denoting the local electric field. Note that Eq. (18) represents the quantity that is estimated by conventional PAT reconstruction algorithms.

Equation (18) reveals that image contrast in RF-based PAT is determined by the electrical conductivity of the material, which is described by the complex permittivity, as well as variations in the illuminating electric field at temporal frequency component $\omega = \omega_c$. Because only the electrical conductivity is intrinsic to the object material, it is desirable to make $|\tilde{E}(\mathbf{r},\omega_c)|^2$ as uniform as possible, so one can unambiguously interpret as the distribution of the conductivity. It has been demonstrated in computer-simulation and experimental studies [40] that estimates of $A(\mathbf{r})$ produced by conventional image reconstruction algorithms can be nonuniform and contain distortions due to diffraction of the electromagnetic wave within the object to be imaged. There remains a need to develop improved image reconstruction methods to mitigate these.

The complex permittivity of tissue has a strong dependence on the water content, temperature, and ion concentration. Because of this, any variations in blood flow in tissue will give rise to changes in the quantity of water and consequently to changes in its complex permittivity. RF-based PAT therefore has the high sensitivity to tissue properties of a microwave technique but requires solution of a tractable acoustic inverse source problem for image reconstruction.

Functional PAT

A highly desirable characteristic of PAT is its ability to provide detailed functional, in additional to anatomical, information regarding biological systems. In this section, we provide a brief review of functional imaging using PAT. For additional details, the reader is referred to parts IX and X in reference [55] and the references therein.

Due to optical contrast mechanism discussed in section "Image Contrast in Laser-Based PAT," laser-based functional PAT operating in the near-infrared (NIR) frequency range can be employed to determine information regarding the oxygenated and deoxygenated hemoglobin within the blood of tissues. This can permit the study of vascularization and hemodynamics, which is relevant to brain imaging and cancer detection.

Functional PAT imaging of hemoglobin can be achieved by exploiting the known characteristic absorption spectra of oxygenated hemoglobin (HbO_2) and deoxygenated hemoglobin (Hb). Consider the situation where the optical fluence $\Phi_s(\mathbf{r})$ is known, and therefore the optical absorption coefficient $\mu_a(\mathbf{r})$ can be determined from the reconstructed absorbed energy density $A(\mathbf{r})$ via Eq. (13). Let $\mu_a^{\lambda_1}(\mathbf{r})$ and $\mu_a^{\lambda_2}(\mathbf{r})$ denote the reconstructed estimates of $\mu_a(\mathbf{r})$ corresponding to the cases where the wavelength of the optical source is set at λ_1 and λ_2. From knowledge of these two estimates, the hemoglobin oxygen saturation distribution, denoted by $SO_2(\mathbf{r})$, is determined as

$$SO_2(\mathbf{r}) = \frac{\mu_a^{\lambda_2}(\mathbf{r})\epsilon_{\text{Hb}}^{\lambda_1} - \mu_a^{\lambda_1}(\mathbf{r})\epsilon_{\text{Hb}}^{\lambda_2}}{\mu_a^{\lambda_1}(\mathbf{r})\epsilon_{\Delta\text{Hb}}^{\lambda_2} - \mu_a^{\lambda_2}(\mathbf{r})\epsilon_{\Delta\text{Hb}}^{\lambda_1}}, \tag{19}$$

where $\epsilon_{\text{Hb}}^{\lambda}$ and $\epsilon_{\text{HbO}_2}^{\lambda}$ denote molar extinction coefficients of Hb and HbO_2, and $\epsilon_{\Delta\text{Hb}}^{\lambda} \equiv \epsilon_{\text{HbO}_2}^{\lambda} - \epsilon_{\text{Hb}}^{\lambda}$. The distribution of the total hemoglobin concentration, denoted by $\text{HbT}(\mathbf{r})$, can be determined as

$$\text{HbT}(\mathbf{r}) = \frac{\mu_a^{\lambda_1}(\mathbf{r})\epsilon_{\Delta\text{Hb}}^{\lambda_2} - \mu_a^{\lambda_2}(\mathbf{r})\epsilon_{\Delta\text{Hb}}^{\lambda_1}}{\epsilon_{\text{Hb}}^{\lambda_1}\epsilon_{\text{HbO}_2}^{\lambda_2} - \epsilon_{\text{Hb}}^{\lambda_2}\epsilon_{\text{HbO}_2}^{\lambda_1}}. \tag{20}$$

An experimental investigation of functional PAT imaging of a rat brain was described in [59]. While in different physiological states, a rat was imaged using laser light at wavelengths 584 and 600 nm to excite the photoacoustic signals. A two-dimensional (2D) scanning geometry was employed, and the estimates of $A(\mathbf{r})$ were reconstructed by use of a backprojection reconstruction algorithm. Subsequently, estimates of $SO_2(\mathbf{r})$ and $\text{HbT}(\mathbf{r})$ were computed and are displayed in Fig. 2.

3 Principles of PAT Image Reconstruction

In the remainder of this chapter, we describe some basic principles that underlie image reconstruction in PAT. We begin by considering the image reconstruction problem in its continuous form. Subsequently, issues related to discrete imaging models that are employed in iterative image reconstruction methods are reviewed.

A schematic of a general PAT imaging geometry is shown in Fig. 3. A short laser or RF pulse is employed to irradiate an object and, as described earlier, the thermoacoustic effect results in the generation of a pressure wavefield $p(\mathbf{r}, t)$. The

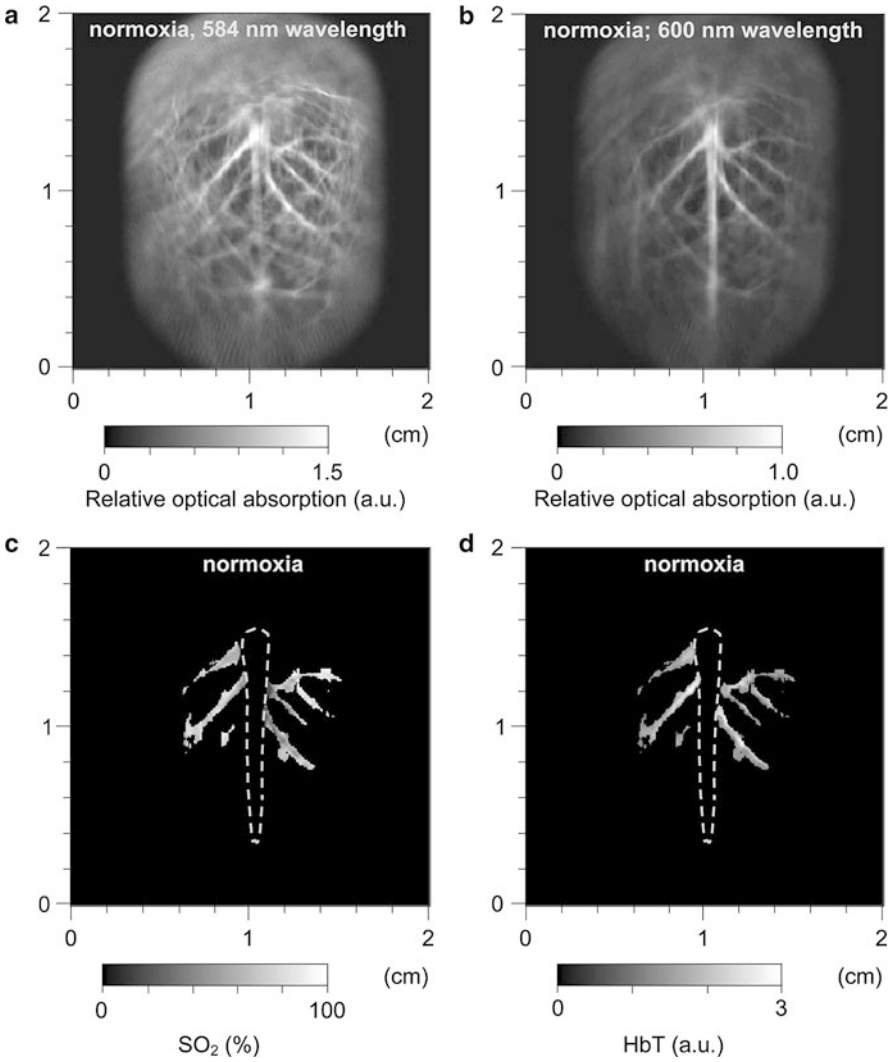


Fig. 2 Noninvasive spectroscopic photoacoustic imaging of HbT and SO_2 in the cerebral cortex of a rat brain. (**a**) and (**b**) Brain images generated by 584 and 600 nm laser light, respectively; (**c**) and (**d**) image of SO_2 and HbT in the areas of the cortical venous vessels (Reproduced from Wang et al. [59])

pressure wavefield propagates out of the object and is measured by use of wideband ultrasonic transducers located on a measurement aperture $\Omega_0 \subset \mathbb{R}^3$, which is a 2D surface that partially or completely surrounds the object. The coordinate $\mathbf{r}_0 \in \Omega_0$ will denote a particular transducer location. Although we will assume that the ultrasound transducers are point-like, it should be noted that alternative

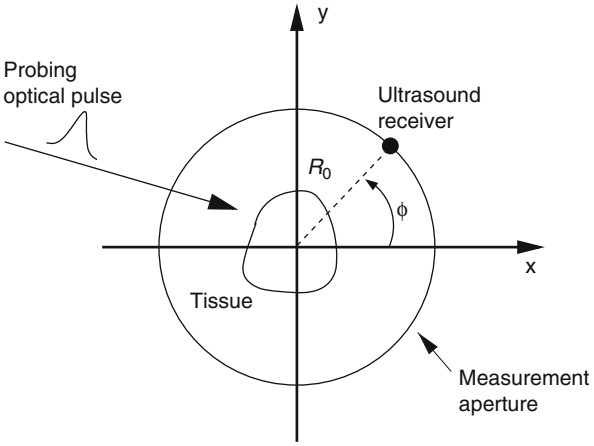


Fig. 3 A schematic of the PAT imaging geometry

implementations of PAT are being actively developed that employ integrating ultrasound detectors [24, 49].

PAT Imaging Models in Their Continuous Forms

When the object possesses homogeneous acoustic properties that match a uniform and lossless background medium, and the duration of the irradiating optical pulse is negligible (acoustic stress confinement is obtained), the pressure wavefield $p(\mathbf{r}_0, t)$ recorded at transducer location $\mathbf{r}_0$ can be expressed [65] as a solution to Eq. (9):

$$p(\mathbf{r}_0, t) = \frac{\beta}{4\pi C_p} \int_V d^3\mathbf{r}\, A(\mathbf{r}) \frac{d}{dt} \frac{\delta\left(t - \frac{|\mathbf{r}_0 - \mathbf{r}|}{c_0}\right)}{|\mathbf{r}_0 - \mathbf{r}|}, \tag{21}$$

where c_0 is the (constant) speed of sound in the object and background medium. The function $A(\mathbf{r})$ is compactly supported, bounded, and nonnegative, and the integration in Eq. (21) is performed over the object's support volume V. Equation (21) represents a canonical imaging model for PAT. The inverse problem in PAT is to determine an estimate of $A(\mathbf{r})$ from knowledge of the measured $p(\mathbf{r}_0, t)$. Note that, as described later, the measured $p(\mathbf{r}_0, t)$ will generally need to be corrected for degradation caused by the temporal and spatial response of the ultrasound transducer.

The imaging model in Eq. (21) can be expressed in an alternate but mathematically equivalent form as

$$g(\mathbf{r}_0, t) = \int_V d^3\mathbf{r}\, A(\mathbf{r})\, \delta\left(t - \frac{|\mathbf{r}_0 - \mathbf{r}|}{c_0}\right), \tag{22}$$

where the integrated data function $g(\mathbf{r}_0, t)$ is defined as

$$g(\mathbf{r}_0, t) \equiv \frac{4\pi C_p c_0}{\beta} t \int_0^t dt' \; p(\mathbf{r}_0, t'). \tag{23}$$

Note that $g(\mathbf{r}_0, t)$ represents a scaled version of the acoustic velocity potential $\phi(\mathbf{r}_0, t)$. Equation (22) represents a spherical Radon transform [22,43] and indicates that the integrated data function describes integrals over concentric spherical surfaces of radii $c_0 t$ that are centered at the receiving transducer location $\mathbf{r}_0$. When these spherical surfaces can be approximated as planes, which would occur when imaging sufficiently small objects that are placed at the center of the scanning system, Eq. (22) can be approximated as a 3D Radon transform [30, 31].

Universal Backprojection Algorithm

A number of analytic image reconstruction algorithms [22, 34, 64, 65] for PAT have been developed in recent years for inversion of Eq. (21) or (22). A detailed description of analytic algorithms will be provided in ▸ Mathematics of Photoacoustic and Thermoacoustic Tomography. However, the so-called universal backprojection algorithm [64] is reviewed below.

The three canonical measurement geometries in PAT employ measurement apertures Ω_0 that are planar [66], cylindrical [67], or spherical [61]. The universal backprojection algorithm proposed by Xu and Wang [64] has been explicitly derived for these geometries. In order to present the algorithm in a general form, let S denote a surface, where $S = \Omega_0$ for the spherical and cylindrical geometries. For the planar geometry, let $S = \Omega_0 + \Omega_0'$, where Ω_0' is a planar surface that is parallel to Ω_0 and the object resides between Ω_0 and Ω_0'.

It has been verified that the initial pressure distribution $p(\mathbf{r}, t = 0) = \Gamma A(\mathbf{r})$ can be mathematically determined from knowledge of the measured $p(\mathbf{r}_0, t)$, $\mathbf{r}_0 \in \Omega_0$, by use of the formula

$$p(\mathbf{r}, t = 0) = \frac{1}{\pi} \int_S dS \int_{-\infty}^{\infty} dk \; \tilde{p}(\mathbf{r}_0, k) \left[\mathbf{n}_0^S \cdot \nabla_0 \tilde{G}_k^{(in)}(\mathbf{r}, \mathbf{r}_0) \right], \tag{24}$$

where $\tilde{p}(\mathbf{r}_0, k)$ denotes the temporal Fourier transform of $p(\mathbf{r}_0, t)$ that is defined with respect to the reduced variable $\bar{t} = c_0 t$ as

$$\tilde{p}(\mathbf{r}_0, k) = \int_{-\infty}^{\infty} d\bar{t} \; p(\mathbf{r}_0, \bar{t}) \exp(ik\bar{t}). \tag{25}$$

Here, $\mathbf{n}_0^S$ denotes the unit vector normal to the surface S pointing toward the source, ∇_0 denotes the gradient operator acting on the variable $\mathbf{r}_0$, and $\tilde{G}_k^{(in)}(\mathbf{r}, \mathbf{r}_0) = \frac{\exp(-ik|\mathbf{r}-\mathbf{r}_0|)}{4\pi|\mathbf{r}-\mathbf{r}_0|}$ is a Green's function of the Helmholtz equation.

Equation (24) can be expressed in the form of a filtered backprojection algorithm as

$$p(\mathbf{r}, t = 0) = \int_{\Sigma_0} d\Sigma_0 \frac{b(\mathbf{r}_0, \bar{t} = |\mathbf{r} - \mathbf{r}_0|)}{\Sigma_0}, \tag{26}$$

where Σ_0 is the solid angle of the whole measurement surface Ω_0 with respect to the reconstruction point inside Ω_0. Note that $\Sigma_0 = 4\pi$ for the spherical and cylindrical geometries, while $\Sigma_0 = 2\pi$ for the planar geometry. The solid angle differential $d\Sigma_0$ is given by

$$d\Sigma_0 = \frac{d\Omega_0}{|\mathbf{r} - \mathbf{r}_0|^2} \frac{\mathbf{n}_0^S \cdot (\mathbf{r} - \mathbf{r}_0)}{|\mathbf{r} - \mathbf{r}_0|}, \tag{27}$$

where $d\Omega_0$ is the differential surface area element on Ω_0. The filtered data function $b(\mathbf{r}_0, \bar{t})$ is related to the measured pressure data as

$$b(\mathbf{r}_0, \bar{t}) = 2p(\mathbf{r}_0, \bar{t}) - 2\bar{t} \frac{\partial p(\mathbf{r}_0, \bar{t})}{\partial \bar{t}}. \tag{28}$$

Equation (26) has a simple interpretation. It states that $p(\mathbf{r}, t = 0)$, or equivalently $A(\mathbf{r})$, can be determined by backprojecting the filtered data function onto a collection of concentric spherical surfaces that are centered at each transducer location $\mathbf{r}_0$.

The Fourier-Shell Identity

Certain insights regarding the spatial resolution of images reconstructed in PAT can be gained by formulating a Fourier domain mapping between the measured pressure data and the Fourier components of $A(\mathbf{r})$ [6]. Below we review a mathematical relationship between the pressure wavefield data function and its normal derivative measured on an arbitrary aperture that encloses the object and the 3D Fourier transform of the optical absorption distribution evaluated on concentric (Ewald) spheres [6]. We have referred to this relationship as a "Fourier-shell identity," which is analogous to the well-known Fourier slice theorem of X-ray tomography.

Consider a measurement aperture Ω_0 that is smooth and closed, but is otherwise arbitrary, and let $\hat{s} \in \mathbf{S}^2$ denote a unit vector on the 3D unit sphere $\mathbf{S}^2$. The 3D spatial Fourier transform of $A(\mathbf{r})$, denoted as $\bar{A}(\nu)$, is defined as

$$\bar{A}(\nu) = \int_V d\mathbf{r}\, A(\mathbf{r})\, \mathrm{e}^{-i\nu \cdot \mathbf{r}}, \tag{29}$$

where the 3D spatial frequency vector $\nu = (\nu_x, \nu_y, \nu_z)$ is the Fourier conjugate of $\mathbf{r}$. It has been demonstrated [6] that

$$\bar{A}(\nu = k\hat{s}) = \frac{iC_p}{k\beta\tilde{I}(k)} \int_{\Omega_0} dS' \left[\hat{n}' \cdot \nabla \tilde{p}\left(\mathbf{r}_0', k\right) + ik\hat{n}' \cdot \hat{s}\ \tilde{p}\left(\mathbf{r}_0', k\right)\right] \mathrm{e}^{-ik\hat{s}\cdot\mathbf{r}_0'}, \tag{30}$$

where $\tilde{p}(\mathbf{r}_0, k)$ is defined in Eq. (25), dS' is the differential surface element on Ω_0, and $\hat{n}'$ is the unit outward normal vector to Ω_0 at the point $\mathbf{r}_0' \in \Omega_0$. Equation (30) has been referred to as the *Fourier-shell identity* of PAT. Because $\hat{s}$ can be chosen to specify any direction, $\bar{A}(\nu = k\hat{s})$ specifies the Fourier components of $A(\mathbf{r})$ that reside on a spherical surface of radius $|k|$, whose center is at the origin. Therefore, Eq. (30) specifies concentric "shells" of Fourier components of $A(\mathbf{r})$ from knowledge of $\tilde{p}(\mathbf{r}_0, k)$ and its derivative along the $\hat{n}'$ direction at each point on the measurement aperture. As reviewed below, this will permit a direct and simple analysis of certain spatial resolution characteristics of PAT.

For a 3D time-harmonic inverse source problem, it is well known [10,13,14] that measurements of the radiated wavefield and its normal derivative on a surface that encloses the source specify the Fourier components of the source function that reside on an Ewald sphere of radius $k = \frac{\omega}{c}$, where ω is the temporal frequency. In PAT, the temporal dependence $I(t)$ of the heating function $H(\mathbf{r}, t)$ is not harmonic and, in general, $\tilde{I}(k) \neq 0$. In the ideal case where $I(t) = \delta(t)$, $\tilde{I}(k) = c$. Consequently, when Eq. (30) is applied to each temporal frequency component k of $\tilde{p}(\mathbf{r}_0, k)$, the entire 3D Fourier domain, with exception of the origin, is determined by the resulting collection of concentric spherical shells. This is possible because of the separable form of the heating function in Eq. (6).

Special Case: Planar Measurement Geometry

The Fourier-shell identity can be used to obtain reconstruction formulas for canonical measurement geometries. For example, consider the case of an infinite planar aperture Ω_0. Specifically, we assume a 3D object is centered at the origin of a Cartesian coordinate system, and the measurement aperture Ω_0 coincides with the plane $y = d > R$, where R is the radius of the object. In this situation, $\mathbf{r}_0' = (x', d, z')$, $dS' = dx'dz'$, and $\hat{n}' = \hat{y}$, where $\hat{y}$ denotes the unit vector along the positive y-axis. The components of the unit vector $\hat{s}$ will be denoted as (s_x, s_y, s_z). Equation (30) can be expressed as the following two terms:

$$\bar{A}(\nu = k\hat{s}) = \bar{A}_1(\nu = k\hat{s}) + \bar{A}_2(\nu = k\hat{s}), \tag{31}$$

where

$$\bar{A}_1(\nu = k\hat{s}) \equiv \frac{iC_p}{kc\beta\tilde{I}(k)} \mathrm{e}^{-ikds_y} \iint_{\infty} dx'dz' \left.\frac{\partial\tilde{p}(x', y, z', k)}{\partial y}\right|_{y=d} \mathrm{e}^{-ik(x's_x + z's_z)}, \tag{32}$$

and

$$\bar{A}_2(\nu = k\hat{s}) \equiv \frac{-C_p\, s_y}{c\beta\tilde{I}(k)} \mathrm{e}^{-ikds_y} \iint_{\infty} dx'dz'\ \tilde{p}(x', d, z', k)\, \mathrm{e}^{-ik(x's_x + z's_z)}, \tag{33}$$

where, without confusion, we employ the notation $\tilde{p}(x, y, z, k) = \tilde{p}(\mathbf{r}_0, k)$.

It can be readily verified that Eqs. (32) and (33) can be reexpressed as

$$\bar{A}_1(\nu = k\hat{s}) \equiv \frac{iC_p}{kc\beta\tilde{I}(k)} \mathrm{e}^{-ikds_y} \frac{\partial}{\partial y} \bar{\tilde{p}}(ks_x, y, ks_z, k)\Big|_{y=d} \tag{34}$$

and

$$\bar{A}_2(\nu = k\hat{s}) \equiv \frac{-C_p s_y}{c\beta\tilde{I}(k)} \mathrm{e}^{-ikds_y} \bar{\tilde{p}}(ks_x, d, ks_z, k), \tag{35}$$

where $\bar{\tilde{p}}(\nu_x, y, \nu_z, k)$ is the 2D spatial Fourier transform of $\tilde{p}(x, y, z, k)$ with respect to x and z (the detector plane coordinates):

$$\bar{\tilde{p}}(\nu_x, y, \nu_z, k) \equiv \frac{1}{4\pi^2} \iint_{\infty} dx dz \; \tilde{p}(x, y, z, k) \, \mathrm{e}^{-i(x\nu_x + z\nu_z)}. \tag{36}$$

The free-space propagator for time-harmonic homogeneous wavefields (see, e.g., Ref. [36], Chapter 4.2) can be utilized to compute the derivative in Eq. (34) as

$$\frac{\partial \bar{\tilde{p}}(ks_x, y, ks_z, k)}{\partial y} = ik\sqrt{1 - s_x^2 - s_z^2} \; \bar{\tilde{p}}(ks_x, y, ks_z, k) = iks_y \; \bar{\tilde{p}}(ks_x, y, ks_z, k), \tag{37}$$

where $s_y \geq 0$. Equations (34)–(37) and (31) establish that

$$\bar{A}(\nu = k\hat{s}) = 2\,\bar{A}_2(\nu = k\hat{s}) \quad \text{for} \quad s_y \geq 0. \tag{38}$$

Equation (38) permits estimation of $\bar{A}(\nu = k\hat{s})$ on concentric half shells in the domain $\nu_y \geq 0$ and is mathematically equivalent to previously studied Fourier-based reconstruction formulas [29, 66]. Note that $A(\mathbf{r})$ is real valued, and therefore the Fourier components in the domain $\nu_y < 0$ can be determined by use of the Hermitian conjugate symmetry property of the Fourier transform.

Spatial Resolution from a Fourier Perspective

The Fourier-shell identity described in section "The Fourier-Shell Identity" is a convenient tool for understanding the spatial resolution characteristics of PAT. Below, we analyze the effects of finite transducer temporal bandwidth and aperture size on spatial resolution [6, 62]. The analysis is applicable to any measurement aperture Ω_0 that corresponds to a coordinate surface of a curvilinear coordinate system.

Effects of Finite Transducer Bandwidth

Consider a point-like ultrasonic transducer whose temporal filtering characteristics are described by the transfer function $\tilde{B}(k;\mathbf{r}_0)$. The $\mathbf{r}_0$-dependence of $\tilde{B}(k;\mathbf{r}_0)$ permits transducers located at different measurement locations to be characterized by distinct transfer functions. The temporal Fourier transform of the measured pressure signal that has been degraded by the temporal response of the transducer will be denoted as $\tilde{p_b}(\mathbf{r}_0,k)$, in order to distinguish it from the ideal pressure signal $\tilde{p}(\mathbf{r}_0,k)$. Because the temporal transducer response can be described by a linear time-invariant system, the degraded and ideal pressure data are related as

$$\tilde{p_b}(\mathbf{r}_0,k) = \tilde{B}(k;\mathbf{r}_0)\tilde{p}(\mathbf{r}_0,k). \tag{39}$$

Consider the case where Ω_0 corresponds to a coordinate surface of a curvilinear coordinate system, that is, $\mathbf{r}_0 \in \Omega_0$ is a vector that varies in only two of its three components. For such surfaces, $B\left(k;\mathbf{r}_0'\right)$ can be interpreted as a 3D function that does not vary in the $\hat{n}'$ direction and therefore $\hat{n}'\cdot\nabla\tilde{B}\left(k;\mathbf{r}_0'\right) = 0$. If the Fourier-shell identity in Eq. (30) is applied with the degraded data function $\tilde{p_b}(\mathbf{r}_0,k)$ replacing the ideal data, the 3D Fourier components of the resulting image, denoted by $A_b(\mathbf{r})$, are recovered as

$$\bar{A_b}(\nu = k\hat{s}) = \frac{iC_p}{k\beta\tilde{I}(k)}\int_{\Omega_0} dS'\,\tilde{B}\left(k;\mathbf{r}_0'\right)\left[\hat{n}'\cdot\nabla\tilde{p}\left(\mathbf{r}_0',k\right) + ik\hat{n}'\cdot\hat{s}\,\tilde{p}\left(\mathbf{r}_0',k\right)\right]e^{-ik\hat{s}\cdot\mathbf{r}_0'}. \tag{40}$$

On comparison of Eqs. (30) and (40), we observe that the spatially variant transducer transfer function $\tilde{B}(k;\mathbf{r}_0)$ modulates the integrand of the Fourier-shell identity. In this general case, the spatial resolution of $A(\mathbf{r})$ will be spatially variant.

If a collection of identical transducers spans Ω_0, $\tilde{B}(k;\mathbf{r}_0) = \tilde{B}(k)$ will not depend on $\mathbf{r}_0$ and Eq. (40) reduces to the simple form

$$\bar{A_b}(\nu = k\hat{s}) = \tilde{B}(k)\bar{A}(\nu = k\hat{s}), \tag{41}$$

where $\bar{A}(\nu = k\hat{s})$ is the exact Fourier data as defined in Eq. (30). As shown in Fig. 4, the one-dimensional (1D) transfer function $\tilde{B}(k)$ of the transducer serves as a radially symmetric 3D filter that modifies $\bar{A}(\nu = k\hat{s})$. This establishes that the image degradation is described by a shift-invariant linear system:

$$A_b(\mathbf{r}) = A(\mathbf{r}) * B(\mathbf{r}), \tag{42}$$

where $*$ denotes a 3D convolution and

$$B(\mathbf{r}) = B(|\mathbf{r}|) = \int_0^\infty dk\,\tilde{B}(k)\frac{\sin(k|\mathbf{r}|)}{k|\mathbf{r}|}\,k^2 \tag{43}$$

is the point-spread function. Equation (43) is consistent with the results derived in Ref. [62].

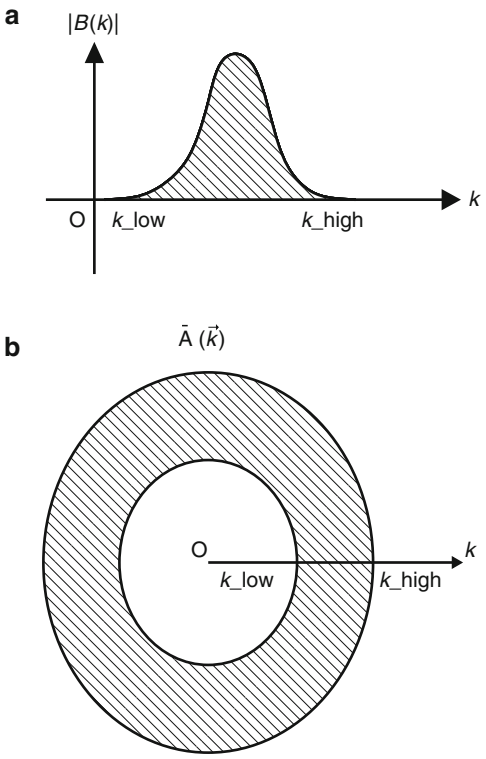


Fig. 4 (**a**) An example of a transducer transfer function $\tilde{B}(k)$. (**b**) The 1D function $\tilde{B}(k)$ acts as a radially symmetric filter in the 3D Fourier domain. The *shaded region* indicates the bandpass of 3D Fourier components that results from application of Eq. (41) (Reproduced from Anastasio et al. [6])

Effects of Nonpoint-Like Transducers

In addition to a nonideal temporal response, a transducer cannot be exactly point-like and will have a finite aperture size. To understand the effects of this on spatial resolution, we consider here a transducer that has an ideal temporal response (i.e., $\tilde{B}(k;\mathbf{r}_0) = 1$) but a finite aperture size [6]. We will assume that the surface of the transducer aperture is a subset of the measurement aperture Ω_0.

It will be useful to employ a local 3D coordinate system whose origin coincides with the center of the detecting surface $\Omega_L \subseteq \Omega_0$, for a transducer at some arbitrary but fixed location $\mathbf{r}_0' \in \Omega_0$. A vector in this system will be denoted as $\mathbf{r}_L$, and the collection of $\mathbf{r}_L \in \Omega_L$ spans all locations on the detecting surface of this transducer. For a transducer located at a different position $\mathbf{r}_0 \in \Omega_0$, the local coordinate vector will be denoted as

$$\mathbf{r}_L^0 = T_{\mathbf{r}_0}\{\mathbf{r}_L\}, \tag{44}$$

where $T_{\mathbf{r}_0}\{\cdot\}$ denotes the corresponding coordinate transformation. This indicates that the collection of vectors $\mathbf{r}_L^0$ corresponding to $\mathbf{r}_L \in \Omega_L$ reside in a local coordinate system whose origin is at $\mathbf{r}_0$ and span all locations on the detecting surface of the transducer centered at that location.

The measured pressure data $\tilde{p_a}(\mathbf{r}_0, k)$, where the subscript "$a$" denotes the data are obtained in the presence of a finite aperture, can be expressed as

$$\tilde{p_a}(\mathbf{r}_0, k) = \int_{\Omega_L} dS_L \, W(\mathbf{r}_L) \tilde{p}\left(\mathbf{r}_0 + \mathbf{r}_L^0, k\right), \tag{45}$$

where dS_L is the differential surface element on Ω_L and the aperture function $W(\mathbf{r}_L)$ describe the sensitivity of the transducers at location $\mathbf{r}_L$ on their surfaces. We assume the aperture function is identical for all transducers, and therefore $W(\mathbf{r}_L)$ can be described simply in terms of the local coordinate $\mathbf{r}_L$. Note that $\mathbf{r}_L^0$ is a function of $\mathbf{r}_L$, as described by Eq. (44).

If the Fourier-shell identity in Eq. (30) is applied with the degraded data function $\tilde{p_a}(\mathbf{r}_0, k)$, the 3D Fourier components of the corresponding image $A_a(\mathbf{r})$ are recovered as

$$\begin{aligned}\bar{A_a}(\nu = k\hat{s}) = &\frac{iC_p}{k\beta\tilde{I}(k)} \int_{\Omega_L} dS_L \, W(\mathbf{r}_L) \\ &\times \int_{\Omega_0} d\Omega_0' \, \mathrm{e}^{-ik\hat{s}\cdot\mathbf{r}_0'} \left[\hat{n}'\cdot\nabla\tilde{p}\left(\mathbf{r}_0' + \mathbf{r}_L^0, k\right) + ik\hat{n}'\cdot\hat{s}\;\tilde{p}\left(\mathbf{r}_0' + \mathbf{r}_L^0, k\right)\right].\end{aligned} \tag{46}$$

By use of the change-of-variable $\mathbf{r}_0 \equiv \mathbf{r}_0' + \mathbf{r}_L^0$ in Eq. (46), one obtains

$$\begin{aligned}\bar{A_a}(\nu = k\hat{s}) = &\frac{iC_p}{k\beta\tilde{I}(k)} \int_{\Omega_L} dS_L \, W(\mathbf{r}_L) \\ &\times \int_{\Omega_0} d\Omega_0 \, \mathrm{e}^{-ik\hat{s}\cdot\left(\mathbf{r}_0 - \mathbf{r}_L^0\right)} [\hat{n}'\cdot\nabla\tilde{p}(\mathbf{r}_0, k) + ik\hat{n}'\cdot\hat{s}\;\tilde{p}(\mathbf{r}_0, k)],\end{aligned} \tag{47}$$

which cannot be simplified further.

The fact that Eq. (47) does not reduce to a simple form analogous to Eq. (41) reflects that the image degradation due to a finite transducer aperture is generally not described by a shift-invariant system [62]. A shift-invariant description is obtained for planar apertures where Eq. (44) reduces to $\mathbf{r}_L^0 = \mathbf{r}_L$, where $\mathbf{r}_L^0$ no longer has a dependence on $\mathbf{r}_0$. In this case, Eq. (47) can be expressed as

$$\bar{A_a}(\nu = k\hat{s}) = \bar{W}(k\hat{s})\bar{A_a}(\nu = k\hat{s}), \tag{48}$$

where

$$\bar{W}(k\hat{s}) \equiv \int_{\Omega_L} dS_L \, W(\mathbf{r}_L) \, \mathrm{e}^{ik\hat{s}\cdot\mathbf{r}_L}. \tag{49}$$

Because, in this case, $\mathbf{r}_L$ resides on a plane and $W(\mathbf{r}_L)$ is a real-valued function, Eq. (49) corresponds to the complex conjugate of the 2D Fourier transform of the aperture function. The point-spread function obtained by computing the 3D inverse Fourier transform of $\bar{W}(k\hat{s})$ reduces to a result given in [62].

4 Speed-of-Sound Heterogeneities and Acoustic Attenuation

In practice, the object to be imaged may not possess uniform acoustic properties, and the images reconstructed by use of algorithms that ignore this can contain artifacts and distortions. Below, we review some methods that can compensate for an object's frequency-dependent acoustic attenuation and heterogeneous speed-of-sound distribution.

Frequency-Dependent Acoustic Attenuation

Because the thermoacoustically inducted pressure signals measured in PAT are broadband and ultrasonic attenuation is frequency dependent, in certain applications it may be important to compensate for this effect. Below, we describe a method described in [37] for achieving this.

Acoustic waves propagating in a lossy medium are attenuated with a linear attenuation coefficient $\alpha(\omega)$ of the general form [53]

$$\alpha(\omega) = \alpha_0 |\omega|^n , \tag{50}$$

where ω is the angular frequency of the wave [53]. For ultrasonic waves in tissue, $n \approx 1$ and $\alpha_0 \approx (10^{-7}/2\pi)$ cm^{-1} rad^{-1} s.

Assuming a uniform speed-of-sound distribution, a photoacoustic wave equation with consideration of acoustic attenuation can be expressed as [37]

$$\nabla^2 p(\mathbf{r}, t) - \frac{1}{c_0^2}\frac{\partial^2}{\partial t^2} p(\mathbf{r}, t) + L(t) * p(\mathbf{r}, t) = -\frac{\beta}{C_p} A(\mathbf{r}) \frac{\partial}{\partial t} I(t) , \tag{51}$$

where $*$ denotes temporal convolution, c_0 is now a reference phase velocity, and the function $L(t)$ describes the effect of acoustic attenuation and is defined as

$$L(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} d\omega \left(K(\omega)^2 - \frac{\omega^2}{c_0^2} \right) \exp(-i\omega t), \tag{52}$$

where

$$K(\omega) \equiv \frac{\omega}{c(\omega)} + i\alpha(\omega). \tag{53}$$

Note that in this section, $p(\mathbf{r}, t)$ denotes the pressure that is affected by acoustic attenuation.

The phase velocity, denoted here by $c(\omega)$, also has a temporal frequency dependence according to the Kramers–Kronig relations. For $n = 1$, this relationship is given by

$$\frac{1}{c(\omega)} = \frac{1}{c_0} - \frac{2}{\pi}\alpha_0 \ln\left|\frac{\omega}{\omega_0}\right|, \tag{54}$$

where ω_0 is the reference frequency for which $c(\omega_0) = c_0$.

Let $\widetilde{p}(\mathbf{r}, \omega)$ denote the temporal Fourier transform of the pressure data:

$$\widetilde{p}(\mathbf{r}, \omega) = \int_{-\infty}^{\infty} dt\ p(\mathbf{r}, t) \exp(i\omega t). \tag{55}$$

It has been shown [37] that the Fourier transform of the attenuated data $\widetilde{p}(\mathbf{r}, \omega)$ is related to the unattenuated data $p_{ideal}(\mathbf{r}, t)$ as

$$\begin{aligned}\widetilde{p}(\mathbf{r}, \omega) = I(\omega) &\left(\frac{c_0}{c(\omega)} + i c_0 \alpha_0 \mathrm{sgn}(\omega)\right)^{-1} \\ &\times \int_{-\infty}^{\infty} p_{ideal}(\mathbf{r}, t) \exp\left\{i\left[\omega \frac{c_0}{c(\omega)} + i c_0 \alpha_0 |\omega|\right] t\right\} dt,\end{aligned} \tag{56}$$

where

$$p_{ideal}(\mathbf{r}, t) = \frac{\beta}{C_p} \int d\mathbf{r}' A(\mathbf{r}') \frac{d}{dt} \frac{\delta\left(t - \frac{|\mathbf{r}-\mathbf{r}'|}{c_0}\right)}{4\pi |\mathbf{r} - \mathbf{r}'|}, \tag{57}$$

is the solution to the photoacoustic wave equation in the absence of attenuation.

Equation (56) permits one to investigate the effect of acoustic attenuation in PAT. It can also be discretized to produce a linear system of equations that can be inverted numerically for removal of the effects of acoustic dispersion in the measured photoacoustic signals. Subsequently, a conventional PAT image reconstruction algorithm could be employed to estimate $A(\mathbf{r})$. A numerical example of this is provided in Ref. [37].

Weak Variations in the Speed-of-Sound Distribution

The conventional PAT imaging models described in section "PAT Imaging Models in Their Continuous Forms" assume that the object's speed of sound is constant and equal to that of the background medium. In certain biomedical imaging applications,

this assumption does not reasonably hold true. For example, the speed of sound of breast tissue can vary from 1,400 to 1,540 m/s. Acoustic inhomogeneities can introduce significant wavefront aberrations in the photoacoustic signal that are not accounted for in the available reconstruction algorithms.

For a weakly acoustic scattering object, with consideration of phase aberrations due to the acoustic heterogeneities effects, the forward PAT imaging model can be expressed as a generalized Radon transform [4, 63]

$$\hat{g}(\mathbf{r}_0, \bar{t}) = \int_V d^3\mathbf{r}\, A(\mathbf{r})\delta[\bar{t} - c_0 t_f(\mathbf{r}, \mathbf{r}_0)]\frac{c_0 t_f(\mathbf{r}, \mathbf{r}_0)}{|\mathbf{r}_0 - \mathbf{r}|}, \tag{58}$$

where $t_f(\mathbf{r}, \mathbf{r}_0)$ is the time of flight (TOF) for a pressure wave to travel from point $\mathbf{r}$ within the object to transducer location $\mathbf{r}_0$. For objects possessing weak acoustic heterogeneities, the TOF can be computed accurately as

$$t_f(\mathbf{r}, \mathbf{r}_0) = \int_{\mathbf{r}' \in L(\mathbf{r}, \mathbf{r}_0)} d^3\mathbf{r}' \quad \frac{1}{c(\mathbf{r}')}, \tag{59}$$

where $c(\mathbf{r})$ is the spatially variant acoustic speed and the set $L(\mathbf{r}, \mathbf{r}_0)$ describes a line connecting $\mathbf{r}_0$ and $\mathbf{r}$.

The generalized Radon transform describes weighted integrals of $A(\mathbf{r})$ over iso-TOF surfaces that are not spherical in general. The iso-TOF surfaces are determined by the heterogeneous acoustic speed distribution $c(\mathbf{r})$ of the object. In the absence of acoustic heterogeneities, these are spherical surfaces with varying radii that are centered at $\mathbf{r}_0$, and Eq. (58) reduces to the spherical Radon transform in Eq. (22). To establish this imaging model for PAT imaging, the iso-TOF surfaces that Eq. (58) integrates over need to be determined explicitly by use of a priori knowledge of the speed-of-sound distribution $c(\mathbf{r})$. Estimates of $c(\mathbf{r})$ can be obtained by performing an adjunct ultrasound computed tomography study of the object [25]. Subsequently, ray-tracing methods can be employed to identify the iso-TOF surfaces for each transducer position $\mathbf{r}_0$. Once these path lengths are computed, the points in the object that have the same path lengths can be grouped together to form iso-TOF surfaces. No known analytic methods are available for inversion of Eq. (58). Accordingly, iterative methods have been employed for image reconstruction [3, 25].

A higher-order geometrical acoustics-based imaging model has also been recently proposed [42] that takes into account the first-order effect in the amplitude of the measured signal and higher-order perturbation to the travel times. By incorporating higher-order approximations to the travel time that incorporates the effect of ray bending, the accuracy of reconstructed images was significantly improved. More general reconstruction methods based on the concept of time-reversal are discussed in chapter ▸ Mathematics of Photoacoustic and Thermoacoustic Tomography.

5 Data Redundancies and the Half-Time Reconstruction Problem

In this section, we review data redundancies that result from symmetries in the PAT imaging model [2, 5, 51, 70], which are related to the so-called half-time reconstruction problem of PAT [4]. Specifically, we describe how an image can be reconstructed accurately from knowledge of half of the temporal components recorded at all transducer locations on a closed measurement aperture.

Data Redundancies

Consider the spherical Radon transform imaging model in Eq. (22). Two half-time data functions $g^{(1)}(\mathbf{r}_0, \bar{t})$ and $g^{(2)}(\mathbf{r}_0, \bar{t})$ can be defined as

$$g^{(1)}(\mathbf{r}_0, \bar{t}) = \begin{cases} g(\mathbf{r}_0, \bar{t}) & : \quad R_0 - R_A \leq \bar{t} \leq R_0 \\ 0 & : \quad \text{otherwise,} \end{cases} \tag{60}$$

and

$$g^{(2)}(\mathbf{r}_0, \bar{t}) = \begin{cases} g(\mathbf{r}_0, \bar{t}) & : \quad R_0 < \bar{t} \leq R_0 + R_A \\ 0 & : \quad \text{otherwise.} \end{cases} \tag{61}$$

Here, R_0 denotes the radius of the measurement aperture Ω_0, and R_A denotes the radius of support of $A(\mathbf{r})$. We assume that the object is acoustically homogeneous with speed of sound c_0 and $\bar{t} \equiv c_0 t$. Note that the data functions $g^{(1)}(\mathbf{r}_0, \bar{t})$ and $g^{(2)}(\mathbf{r}_0, \bar{t})$ each cover different halves of the complete data domain $\Omega_0 \times [R_0 - R_A, R_0 + R_A]$, and therefore $g(\mathbf{r}_0, \bar{t}) = g^{(1)}(\mathbf{r}_0, \bar{t}) + g^{(2)}(\mathbf{r}_0, \bar{t})$.

In the limit where $R_0 \to \infty$, the spherical Radon transform reduces to a conventional Radon transform that integrates over 2D planes. In that case, an obvious conjugate-view symmetry exists [10], and therefore, either of the half-time data functions $g^{(1)}(\mathbf{r}_0, \bar{t})$ and $g^{(2)}(\mathbf{r}_0, \bar{t})$ contains enough information, in a mathematical sense, for exact image reconstruction. Accordingly, a twofold data redundancy exists because the complete data function $g(\mathbf{r}_0, \bar{t})$ contains twice as much information as is theoretically necessary for exact image reconstruction.

In the case where R_0 is finite, a simple conjugate view symmetry does not exist. Nevertheless, it has been demonstrated that a twofold data redundancy exists in the complete data function $g(\mathbf{r}_0, \bar{t})$. This has been heuristically [50] and mathematically [5] by use of a layer-stripping procedure [2,5,50,70]. This established that $A(\mathbf{r})$ can be recovered uniquely and stably from knowledge of either half-time data functions $g^{(1)}(\mathbf{r}_0, \bar{t})$ or $g^{(2)}(\mathbf{r}_0, \bar{t})$. A similar conclusion has been derived in Ref. [22] using a different mathematical approach.

Analytic inversion formulae for recovering $A(\mathbf{r})$ from knowledge of the half-time data functions $g^{(1)}(\mathbf{r}_0, \bar{t})$ or $g^{(2)}(\mathbf{r}_0, \bar{t})$ are not currently available. However, iterative reconstruction algorithms can be employed [4, 5] to determine $A(\mathbf{r})$.

Mitigation of Image Artifacts Due to Acoustic Heterogeneities

If the spatially variant speed of sound $c(\mathbf{r})$ is known, one can numerically invert a discretized version of Eq. (58) to determine an estimate of $A(\mathbf{r})$ [3]. However, in many applications of PAT, $c(\mathbf{r})$ is not known, and images are simply reconstructed by use of algorithms that assume a constant speed of sound. This can result in conspicuous image artifacts.

Let $\hat{g}(\mathbf{r}_0, \bar{t})$ denote a data function that is contaminated by the effects of speed-of-sound variations within the object that is related to $A(\mathbf{r})$ according to Eq. (58). Let $\hat{A}(\mathbf{r})$ denote an estimate of $A(\mathbf{r})$ that is reconstructed from $\hat{g}(\mathbf{r}_0, \bar{t})$ by use of a conventional reconstruction algorithm that assumes an acoustically homogeneous object. The quantities $\hat{g}^{(1)}(\mathbf{r}_0, \bar{t})$ and $\hat{g}^{(2)}(\mathbf{r}_0, \bar{t})$ denote half-time data functions that are defined in analogy with Eqs. (60) and (61) with $g(\mathbf{r}_0, \bar{t})$ replaced by $\hat{g}(\mathbf{r}_0, \bar{t})$.

An image reconstructed from $\hat{g}^{(1)}(\mathbf{r}_0, \bar{t})$ can sometimes contain reduced artifact levels as compared to one reconstructed from the complete data $\hat{g}(\mathbf{r}_0, \bar{t})$. To demonstrate this, in the discussion below, we consider the 2D problem and the spatially variant speed-of-sound distribution shown in Fig. 5. This speed-of-sound distribution is comprised of two uniform concentric disks that have c_0 and c_1, with $c_0 \neq c_1$, and radii r_0 and r_1, respectively. The background medium is assumed to have a speed of sound, c_0.

The acoustic heterogeneity will cause the data function $\hat{g}(\mathbf{r}_0, \bar{t})$ to differ from the ideal one $g(\mathbf{r}_0, \bar{t})$. The magnitude of this difference will be smaller, in general, for small values of $\bar{t}$ than for large values of $\bar{t}$. This can be understood by noting that, in general, $\left| t_f(\mathbf{r}, \mathbf{r}_0) - \frac{|\mathbf{r}_0 - \mathbf{r}|}{c_0} \right|$ will become larger as the path length through the speed-of-sound heterogeneity increases. This causes the surfaces of integration that contribute to $\hat{g}(\mathbf{r}_0, \bar{t})$ to become less spherical for larger values of $\bar{t}$. Accordingly, the

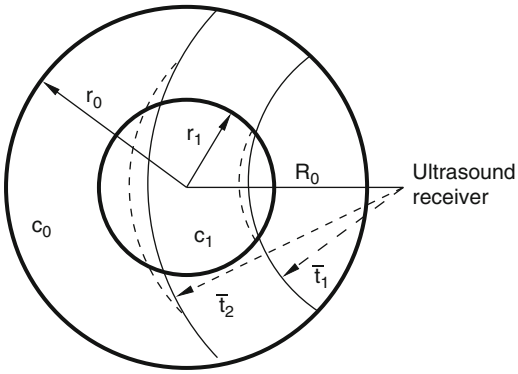


Fig. 5 A speed-of-sound distribution comprised of two uniform concentric regions. Superimposed on the figure are examples of how the surfaces of integration that contribution to the data function $g(\mathbf{r}_0, \bar{t})$ are perturbed

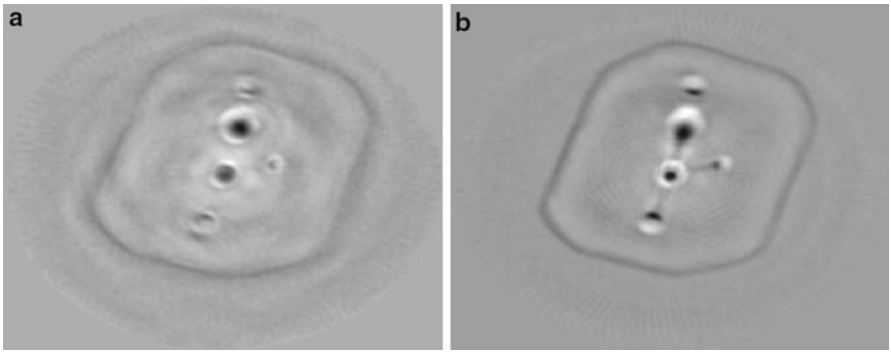


Fig. 6 Images of a phantom object reconstructed from experimentally measured (**a**) full-time, (**b**) first half-time data functions (Reproduced from Anastasio et al. [4])

data function $\hat{g}(\mathbf{r}_0, \bar{t})$ becomes less consistent with the spherical Radon transform model.

The discussion above suggests that a half-time reconstruction method that employs $\hat{g}^{(1)}(\mathbf{r}_0, \bar{t})$ can produce images with reduced artifact and distortion levels than contained in images reconstructed from the complete, or full-time, data $\hat{g}(\mathbf{r}_0, \bar{t})$. An example of this is shown in Fig. 6. The data corresponded to a physical phantom study using a microwave source, as described in Ref. [61]. Measurements were taken at 160 equally spaced positions on the 2D circular scanning aperture of radius 70 mm, and for each measurement, the received pressure signal was sampled at 2,000 points, at a sampling frequency of 50 MHz. Images were reconstructed from full- and half-time data, via the EM algorithm as described in Ref. [4]. The contrast and resolution of the images reconstructed from half-time data appears to be superior to that of the images reconstructed from the full-time data.

6 Discrete Imaging Models

The imaging models discussed so far were expressed in their continuous forms. In practice, PAT imaging systems record temporal and spatial samples of $p(\mathbf{r}_0, t)$, while the absorbed energy density is described by the function $A(\mathbf{r})$. Accordingly, a realistic imaging model should be described mathematically as a continuous-to-discrete (C–D) mapping [8]. Moreover, when iterative reconstruction algorithms are employed, a discrete representation of $A(\mathbf{r})$ is required to establish a suitable discrete-to-discrete approximate imaging model. In this section, we review these concepts within the context of PAT.

The remainder of this section is organized as follows. In section "Continuous-to-Discrete Imaging Models," we review the C–D versions of the continuous-to-continuous (C–C) models in Eqs. (21) and (22). Finite-dimensional object representations are surveyed in section "Finite-Dimensional Object Representations" that

are used to establish the discrete-to-discrete (D–D) models in section "Discrete-to-Discrete Imaging Models." In section "Iterative Image Reconstruction," we briefly review some approaches to iterative image reconstruction that have been applied in PAT. The section concludes with a numerical example that demonstrates the effects of object representation error on image reconstruction accuracy.

Continuous-to-Discrete Imaging Models

In practice, $p(\mathbf{r}_0, t)$ and $g(\mathbf{r}_0, t)$ are discretized temporally and determined at a finite number of receiver locations. The vectors $\mathbf{p}, \mathbf{g} \in \mathbb{R}^N$ will represent lexicographically ordered representations of the sampled data functions, where the dimension N is defined by the product of the number of temporal samples acquired at each transducer location (S) and the number of transducer locations (M). Let Eqs. (21) and (22) be expressed in operator notation as

$$p(\mathbf{r}_0, t) = \mathcal{H}_p A(\mathbf{r}), \tag{62}$$

and

$$g(\mathbf{r}_0, t) = \mathcal{H}_g A(\mathbf{r}). \tag{63}$$

In general, a C–D operator can be interpreted as a discretization operator $\mathcal{D}_{\sigma\tau}$ acting on C–C operator $\mathcal{H}_{CC}$ [8]. Let $\mathbf{y}$ denote $\mathbf{p}$ or $\mathbf{g}$ and let $\mathcal{H}_{CC}$ denote $\mathcal{H}_p$ or $\mathcal{H}_g$. The notation $y_{[n]}$ will be used to denote the nth element of the vector $\mathbf{y}$. The C–D versions of Eqs. (21) and (22) can be expressed as

$$\mathbf{y} = \mathcal{D}_{\sigma\tau}\mathcal{H}_{CC} A(\mathbf{r}) = \mathcal{H}_{CD} A(\mathbf{r}), \tag{64}$$

where $\mathcal{D}_{\sigma\tau}$ is discretization operator that characterizes the temporal and spatial sampling characteristics of the ultrasonic transducer.

For the case where $\mathbf{y} = \mathbf{p}$ and $p(\mathbf{r}_0, t) = \mathcal{H}_p A(\mathbf{r})$, $\mathcal{D}_{\sigma\tau}$ will be denoted as $\mathcal{D}_{\sigma\tau}^{(p)}$ and is defined as

$$p_{[mS+s]} = \left[\mathcal{D}_{\sigma\tau}^{(p)} p(\mathbf{r}_0, t)\right]_{[mS+s]} \equiv \int_{-\infty}^{\infty} dt\ \tau_s(t) \int_{\Omega_0} d\Omega_0\ p(\mathbf{r}_0, t)\sigma_m(\mathbf{r}_0), \tag{65}$$

where $m = 1, 2, \ldots, M$ is the index that specifies the mth transducer location $\mathbf{r}_{0,m}$ on the measurement aperture Ω_0, $s = 1, 2, \ldots, S$ is the index of the time sample, and $\sigma_m(\mathbf{r}_0)$ and $\tau_s(t)$ are functions that describe the spatial and temporal *sampling apertures*, respectively. They are determined by the sampling properties of ultrasonic transducers. In the ideal case, where both apertures are described by Dirac delta functions, the sth temporal sample for the mth transducer location represents the pressure at time $s\Delta T$ and location $\mathbf{r}_{0,m}$, where ΔT is the temporal sampling interval, that is,

$$p_{[mS+s]} = p(\mathbf{r}_{0,m}, s\Delta T). \tag{66}$$

We can express explicitly the C–D imaging model involving the pressure data as

$$p_{[mS+s]} = \int_V d^3\mathbf{r}\ A(\mathbf{r})\, h_{mS+s}(\mathbf{r}), \tag{67}$$

where V denotes the support volume of $A(\mathbf{r})$ and

$$h_{mS+s}(\mathbf{r}) \equiv \int_{-\infty}^{\infty} dt_0\ \tau_s(t_0) \int_{\Omega_0} d\Omega_0\ h(\mathbf{r}, \mathbf{r}_0; t_0)\sigma_m(\mathbf{r}_0) \tag{68}$$

defines a point response function. The kernel $h(\mathbf{r}, \mathbf{r}_0; t_0)$ is defined as

$$h(\mathbf{r}, \mathbf{r}_0; t_0) = \int_{-\infty}^{\infty} dt\ I(t) G(\mathbf{r}, \mathbf{r}_0; t, t_0), \tag{69}$$

where $I(t)$ is the temporal illumination function and $G(\mathbf{r}, \mathbf{r}_0; t, t_0)$ is the Green's function

$$G(\mathbf{r}, \mathbf{r}_0; t, t_0) = \frac{\beta}{4\pi C_p |\mathbf{r} - \mathbf{r}_0|} \frac{d\delta(t)}{dt}\bigg|_{t=t_0-\frac{|\mathbf{r}-\mathbf{r}_0|}{c_0}}. \tag{70}$$

By use of the singular value decomposition of the C–D operator in Eq. (67), a pseudoinverse solution can be computed numerically to estimate $A(\mathbf{r})$ [8].

In order to establish a C–D imaging model involving the integrated pressure data, to first order, we can approximate the integral operator in Eq. (23) as

$$g_{[mS+s]} = \frac{4\pi C_p c_0 s\Delta T}{\beta} \sum_{q=1}^{s} p_{[mS+q]}. \tag{71}$$

For the case where $\mathbf{y} = \mathbf{g}$ and $g(\mathbf{r}_0, t) = \mathcal{H}_g A(\mathbf{r})$, $\mathcal{D}_{\sigma\tau}$ will be denoted as $\mathcal{D}^{(g)}_{\sigma\tau}$ and is defined as

$$\begin{aligned} g_{[mS+s]} &= \left[\mathcal{D}^{(g)}_{\sigma\tau} g(\mathbf{r}_0, t)\right]_{[mS+s]} \\ &\equiv s\Delta T \sum_{q=1}^{s} \int_{-\infty}^{\infty} dt\ \tau_q(t) \int_{\Omega_0} d\Omega_0\ \sigma_m(\mathbf{r}_0) \frac{d}{dt}\left(\frac{g(\mathbf{r}_0, t)}{t}\right). \end{aligned} \tag{72}$$

Note that, in practice, $\mathbf{g}$ is not measured and is computed from the measured $\mathbf{p}$ by use of Eq. (71). Therefore, it is not physically meaningful to interpret $\mathbf{g}$ as being directly sampled from the raw measurement data.

Finite-Dimensional Object Representations

When iterative image reconstruction algorithms are employed, a finite-dimensional representation of $A(\mathbf{r})$ [8] is required. In this section, we review some finite-dimensional representations that have been employed in PAT. In the subsequent section, computer-simulation studies are conducted to demonstrate the effects of error in the object representation.

An N-dimensional representation of $A(\mathbf{r})$ can be described as

$$A_a(\mathbf{r}) = \sum_{n=1}^{N} \theta_{[n]} \phi_n(\mathbf{r}), \tag{73}$$

where the subscript a indicates that $A_a(\mathbf{r})$ is an approximation of $A(\mathbf{r})$. The functions $\phi_n(\mathbf{r})$ are called expansion functions and the expansion coefficients $\theta_{[n]}$ are elements of the N-dimensional vector $\boldsymbol{\theta}$. The goal of iterative image reconstruction methods is to estimate $\boldsymbol{\theta}$, for a fixed choice of the expansion functions $\phi_n(\mathbf{r})$.

The most commonly employed expansion functions are simple image voxels

$$\phi_n(x, y, z) = \begin{cases} 1, & \text{if} \quad |x - x_n|, |y - y_n|, |z - z_n| \leq \epsilon/2 \\ 0, & \text{otherwise} \end{cases} \tag{74}$$

where $\mathbf{r}_n = (x_n, y_n, z_n)$ specify the coordinates of the nth grid point of a uniform Cartesian lattice and ϵ defines the spacing between lattice points.

In PAT, spherical expansion functions of the form

$$\phi_n(x, y, z) = \begin{cases} 1, \text{ if} & \sqrt{(x - x_n)^2 + (y - y_n)^2 + (z - z_n)^2} \leq \epsilon/2 \\ 0, \text{ otherwise} \end{cases} \tag{75}$$

have also proven to be useful [18, 27]. The merit of this kind of expansion function is that the acoustic wave generated by each voxel can be calculated analytically. This facilitates determination of the system matrix utilized by iterative image reconstruction methods, as discussed below. Numerous other effective choices for the expansion functions [38] exist, including wavelets or other sets of functions that can yield sparse object representations [52].

In addition to an infinite number of choices for the expansion functions, there are an infinite number of ways to define the expansion coefficients $\boldsymbol{\theta}$. Some common choices include

$$\theta_{[n]} = \frac{V_{\text{cube}}}{V_{\text{voxel}}} \int_V d^3\mathbf{r}\, \phi_n(\mathbf{r}) A(\mathbf{r}), \tag{76}$$

or

$$\theta_{[n]} = \int_V d^3\mathbf{r}\, \delta(\mathbf{r} - \mathbf{r}_n) A(\mathbf{r}). \tag{77}$$

For a given N, different choices of ϕ_n and $\boldsymbol{\theta}$ will yield object representations that possess different representation errors

$$\delta A(\mathbf{r}) = A(\mathbf{r}) - A_a(\mathbf{r}). \tag{78}$$

An example of the effects of such representation errors on iterative reconstruction methods is provided in section "Iterative Image Reconstruction."

Discrete-to-Discrete Imaging Models

Discrete-to-discrete (D–D) imaging models are required for iterative image reconstruction. These can be obtained systematically by substitution of a finite-dimensional object representation into the C–D imaging model in Eq. (64):

$$\mathbf{y}_a = \mathcal{H}_{\mathrm{CD}}\mathbf{A}_a(\mathbf{r}) = \sum_{n=1}^{N} \theta_{[n]} \mathcal{H}_{\mathrm{CD}}\{\phi_n(\mathbf{r})\} \equiv \mathbf{H}\boldsymbol{\theta}, \tag{79}$$

where the D–D operator $\mathbf{H}$ is commonly referred to as the system matrix. The system matrix $\mathbf{H}$ is of dimension $(MS) \times N$, and an element of $\mathbf{H}$ will be denoted by $H_{[n,m]}$. Note that the data vector $\mathbf{y}_a \neq \mathbf{y}$, due to the fact that a finite-dimensional approximate object representation was employed. In other words, $\mathbf{y}_a$ represents an approximation of the measured pressure data, denoted by $\mathbf{p}_a$, or the corresponding approximate integrated pressure data $\mathbf{g}_a$.

For the case where $\mathbf{y}_a = \mathbf{p}_a$, the system matrix $\mathbf{H}$ will be denoted as $\mathbf{H}^{(p)}$ and its elements are defined as

$$H^{(p)}_{[mS+s,n]} = \int_V d^3\mathbf{r}\, \phi_n(\mathbf{r}) h_{mS+s}(\mathbf{r}) = \mathcal{D}^{(p)}_{\sigma\tau}\{p_n(\mathbf{r}_0, t_0)\}, \tag{80}$$

where $h_{mS+s}(\mathbf{r})$ is defined in Eq. (68) and

$$p_n(\mathbf{r}_0, t_0) = \int_V d^3\mathbf{r}\, \phi_n(\mathbf{r}) h(\mathbf{r}, \mathbf{r}_0; t_0). \tag{81}$$

Equation (80) provides a clear two-step procedure for computing the system matrix. First, $p_n(\mathbf{r}_0, t_0)$ is computed. Physically, this represents the pressure data, in its continuous form, received by an ideal point transducer when the absorbing object corresponds to $\phi_n(\mathbf{r})$. Secondly, a discretization operator is applied that samples the ideal data and degrades it by the transducer response. Alternatively, the elements of the system matrix can be measured experimentally by scanning an object whose form matches the expansion functions through the object volume and recording the resulting pressure signal at each transducer location $\mathbf{r}_{0,m}$, for each value of n (location of expansion function), at time intervals $s\Delta T$. For the case of spherical expansion elements, this approach was implemented in [18].

This two-step approach for determining $\mathbf{H}$ be formulated as

$$\mathbf{H} = \mathbf{S} \circ \mathbf{H}_0, \tag{82}$$

where "$\circ$" denotes an element-wise product. Each element of $\mathbf{H}_0$ is defined as

$$H_{0[mS+s,n]} = p_n(\mathbf{r}_{0,m}, s\Delta T). \tag{83}$$

The $MS \times N$ matrix $\mathbf{S}$ can be interpreted as a sensitivity map, whose elements are defined as

$$S_{[mS+s,n]} = \frac{\mathcal{D}_{\sigma\tau}\{p_n(\mathbf{r}_0, t_0)\}}{p_n(\mathbf{r}_{0,m}, s\Delta T)}. \tag{84}$$

For the case where $\mathbf{y}_a = \mathbf{g}_a$, similar interpretations hold. The system matrix $\mathbf{H}$ will be denoted as $\mathbf{H}^{(g)}$, and its elements are defined as

$$H^{(g)}_{[mS+s,n]} = \mathcal{D}^{(g)}_{\sigma\tau}\{g_n(\mathbf{r}_0, t_0)\}, \tag{85}$$

where

$$g_n(\mathbf{r}_0, t_0) = \frac{4\pi C_p c_0 t_0}{\beta} \int_0^{t_0} d\zeta_0 \int_V d^3\mathbf{r}\, \phi_n(\mathbf{r}) h(\mathbf{r}, \mathbf{r}_0; \zeta_0). \tag{86}$$

Numerical Example: Impact of Representation Error on Computed Pressure Data

Consider a uniform sphere of radius $R_s = 5$ mm as the optical absorber (acoustic source). Assuming Dirac delta (i.e., ideal) temporal and spatial sampling, the pressure data were computed at a measurement location $\mathbf{r}_0$ 65 mm away from the center of the sphere by use of D–D and C–C imaging models. For the uniform sphere, the pressure waveform can be computed analytically as

$$\begin{aligned} p(\mathbf{r}_0, s\Delta T) &= \frac{d}{dt}\left[\frac{\beta}{4\pi C_p c_0 t} g(\mathbf{r}_0, t)\right]\Bigg|_{t=s\Delta T} \\ &= \begin{cases} \frac{\beta c_0^2}{2C_p|\mathbf{r}_0 - \mathbf{r}_c|}(|\mathbf{r}_0 - \mathbf{r}_c| - c_0 s\Delta T), & \text{if } \left|c_0 s\Delta T - |\mathbf{r}_0 - \mathbf{r}|\right| \le R_s \\ 0, & \text{otherwise} \end{cases} \end{aligned} \tag{87}$$

where $\mathbf{r}_c$ is the center of the spherical source and ΔT is the sampling interval. As discussed in section "The Thermoacoustic Effect and Signal Generation," the pressure possesses an "N" shape waveform as shown as the dashed red curve in Fig. 7. Finite-dimensional object representations of the object were obtained according to Eq. (73) with $\phi_n(\mathbf{r})$ corresponding to the uniform spheres described

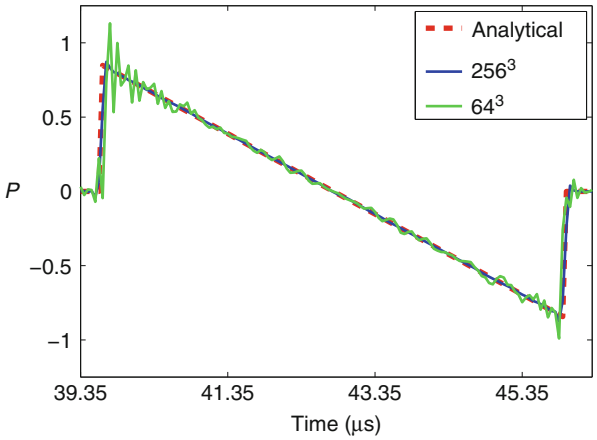


Fig. 7 Pressure data generated by continuous imaging model (*red dash*) and discrete imaging model using 256 × 256 × 256 voxels (*blue solid*) and 64 × 64 × 64 voxels (*green solid*)

in Eq. (75). The expansion coefficients were computed according to Eq. (76). Two approximate object representations were considered. The first representation employed $N = 256^3$ spherical expansion functions of radius 0.04 mm, while the second employed $N = 64^3$ expansion functions of radius 0.16 mm. The resulting pressure signals are shown as Fig. 7, where the speed of sound $c_0 = 1.521\,\text{mm}/\mu\text{s}$ and $\Delta T = 0.05\,\mu\text{s}$. As expected, the error in the computed pressure data increases as the voxel size is increased. In practice, this error would represent a data inconsistency between the measured data and the assumed D–D imaging model, which can result in image artifacts as demonstrated by the example below.

Iterative Image Reconstruction

Once the system matrix $\mathbf{H}$ is determined, as described in the previous section, an estimate of $A(\mathbf{r})$ can be computed in two distinct steps. First, from knowledge of the measured data and system matrix, Eq. (79) is inverted to estimate the expansion coefficients $\boldsymbol{\theta}$. Second, the estimated expansion coefficients are employed with Eq. (73) to determine the finite-dimensional approximation $A_a(\mathbf{r})$. Each of steps introduces error into the final estimate of $A(\mathbf{r})$. In the first step, due to noise in the measured data $\mathbf{y}_a$, modeling errors in $\mathbf{H}$, and/or if $\mathbf{H}$ is not full rank, the true values coefficients $\boldsymbol{\theta}$ cannot generally be determined. The estimated $\boldsymbol{\theta}$ will therefore depend on the definition of the approximate solution and the particular numerical algorithm used to determine it. Even if $\boldsymbol{\theta}$ could somehow be determined exactly, the second step would introduce error due to the approximate finite-dimensional representation of $A(\mathbf{r})$ employed. This error is influenced by the choice of N and $\phi_n(\mathbf{r})$ and is object dependent.

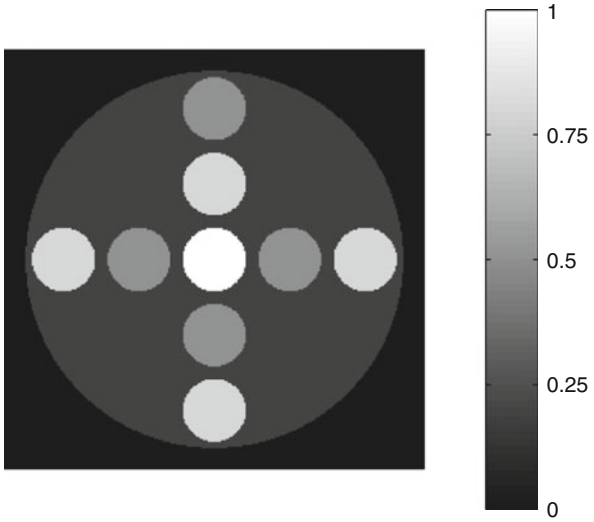


Fig. 8 The 2D numerical phantom $\boldsymbol{\theta}$ representing the object function $A(\mathbf{r})$

Due to the large size of $\mathbf{H}$, iterative methods are often employed to estimate $\boldsymbol{\theta}$. Iterative approaches offer a fundamental and flexible way to incorporate a prior information regarding the object, to improve the accuracy of the estimated $\boldsymbol{\theta}$. A vast literature on iterative image reconstruction methods exists [7, 20, 21, 60], which we leave to the reader to explore. Examples of applications of iterative reconstruction methods in PAT are described in references [1, 4, 5, 18, 48, 69]. A numerical example demonstrating how object representation error can affect the accuracy of iterative image reconstruction is provided next.

Numerical Example: Influence of Representation Error on Image Accuracy

We assume focused transducers are employed that receive only acoustic pressure signals transmitted from the imaging plane and therefore the 3D spherical Radon transform imaging model to a 2D circular mean model. A 2D phantom comprised of uniform disks possessing different gray levels, radii, and locations was assumed to represent $A(\mathbf{r})$. The radius of the phantom was 1.0 (arbitrary units). A finite-dimensional representation $A_a(\mathbf{r})$ was formed according to Eq. (73), with $N = 256^2$ and $\phi_n(\mathbf{r})$ chosen to be conventional pixels described by a 2D version of Eq. (74). The expansion coefficients $\theta_{[n]}$ were computed by use of Eq. (77). Figure 8 displays the computed expansion coefficient vector $\boldsymbol{\theta}$ that has been reshaped into a 256×256 for display purposes.

A circular measurement aperture Ω_0 of radius 1.2 that enclosed the object was employed. At each of 360 uniformly, spaced transducer locations, $\mathbf{r}_{0,m}$, on the measurement circle, simulated pressure data $\mathbf{p}_a$ were computed from the integrated data $\mathbf{g}$ by use of the formula

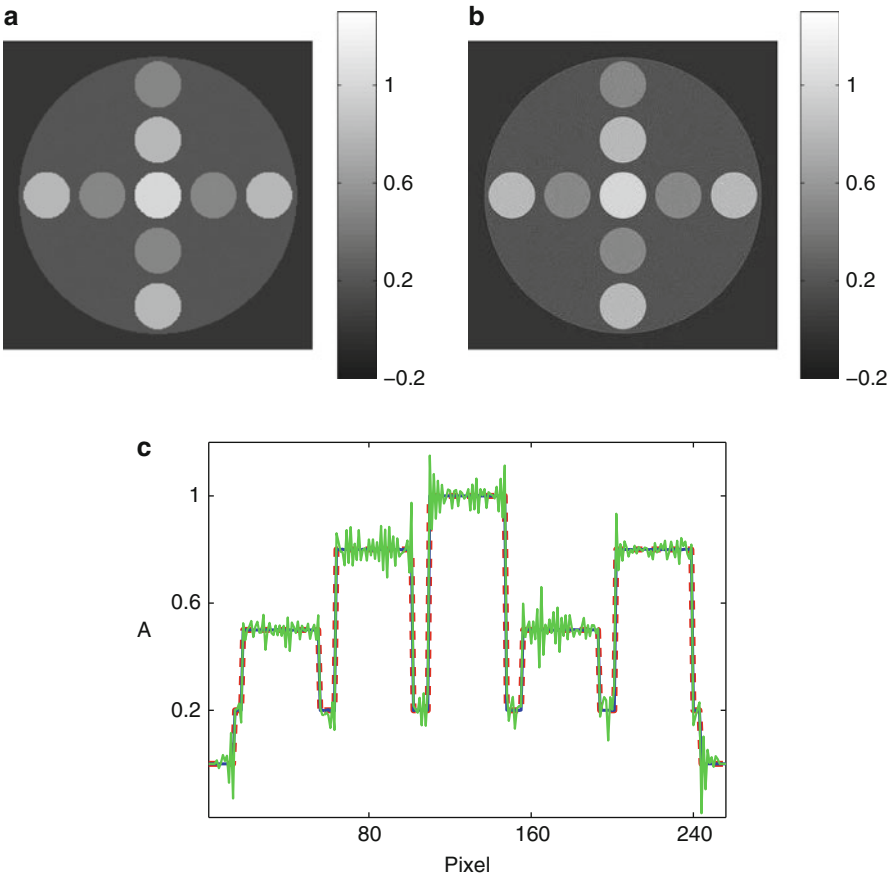


Fig. 9 Images reconstructed by the least-squares conjugate gradient algorithm from pressure data obtained by (**a**) numerical imaging model and (**b**) analytical imaging model. (**c**) Vertical profiles through the center of subfigure (**a**) (*solid blue*), subfigure (**b**) (*solid green*), and Fig. 8 (*dashed red*)

$$p_{a[mS+s]} = \frac{\beta}{4\pi C_p c_0}\left[\frac{g_{[mS+s+1]}/(s+1) - g_{[mS+s-1]}/(s-1)}{2\Delta T^2}\right]. \tag{88}$$

Two versions of the pressure data were computed, corresponding to the cases where $\mathbf{g}$ was computed analytically or by use of the assumed D–D imaging model. These simulated pressure data are denoted by $\mathbf{p}_a^{analy}$ and $\mathbf{p}_a^{num}$, respectively. At each transducer location, 300 temporal samples of $p(\mathbf{r}_0, t)$ were computed. Accordingly, the pressure vector $\mathbf{p}_a$ was a column vector of length 360×300.

The conjugate gradient algorithm was employed to find the least-squares estimate $\hat{\boldsymbol{\theta}}$,

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \|\mathbf{p}_a - \mathbf{H}\boldsymbol{\theta}\|^2, \tag{89}$$

where $\mathbf{p}_a = \mathbf{p}_a^{analy}$ or $\mathbf{p}_a^{num}$. For the noiseless data, the images reconstructed from $\mathbf{p}_a^{analy}$ and $\mathbf{p}_a^{num}$ after 150 iterations are shown as Fig. 9a, b, respectively. The image reconstructed from the data $\mathbf{p}_a^{num}$ is free of significant artifacts and is nearly identical to the original object. This is expected because the finite-dimensional object representation was used to produce the simulated measurement data and establish the system matrix, and therefore the system of equations in Eq. (79) is consistent. Generating simulation data in this way would constitute an "inverse crime." Conversely, the image reconstructed from the data $\mathbf{p}_a^{analy}$ contained high-frequency artifacts due to the fact that the system of equations in Eq. (79) is inconsistent. The error in the reconstructed images could be minimized by increasing the dimension of the approximate object representation. This simple example demonstrates the importance of carefully choosing a finite-dimensional object representation in iterative image reconstruction.

7 Conclusion

Photoacoustic tomography is a rapidly emerging biomedical imaging modality that possesses many challenges for image reconstruction. In this chapter, we have reviewed the physical principles of PAT. Contrast mechanisms in PAT were discussed, and the imaging models that relate the measured photoacoustic wavefields to the sought-after optical absorption distribution were described in their continuous and discrete forms.

Cross-References

- ▸ Iterative Solution Methods
- ▸ Linear Inverse Problems
- ▸ Optical Imaging
- ▸ Tomography

References

1. Anastasio, M.A., Zhang, J.: Image reconstruction in photoacoustic tomography with truncated cylindrical measurement apertures. In: Proceedings of the SPIE Conference, vol. 6086, p. 36 (2006)
2. Anastasio, M.A., Zou, Y., Pan, X.: Reflectivity tomography using temporally truncated data. In: IEEE EMBS/BMES Conference Proceedings, Houston, vol. 2, pp. 921–922. IEEE (2002)
3. Anastasio, M.A., Zhang, J., Pan, X.: Image reconstruction in thermoacoustic tomography with compensation for acoustic heterogeneties. In: Proceedings of the SPIE Medical Imaging Conference, San Diego, vol. 5750, pp. 298–304 (2005)
4. Anastasio, M.A., Zhang, J., Pan, X., Zou, Y., Keng, G., Wang, L.V.: Half-time image reconstruction in thermoacoustic tomography. IEEE Trans. Med. Imaging **24**, 199–210 (2005)
5. Anastasio, M.A., Zhang, J., Sidky, E.Y., Zou, Y., Xia, D., Pan, X.: Feasibility of half-data image reconstruction in 3D reflectivity tomography with a spherical aperture. IEEE Trans. Med. Imaging **24**, 1100–1112 (2005)

6. Anastasio, M.A., Zhang, J., Modgil, D., La Riviere, P.J.: Application of inverse source concepts to photoacoustic tomography. Inverse Probl. **23**(6), S21–S35 (2007)
7. Axelsson, O.: Iterative Solution Methods. Cambridge University Press, Cambridge (1994)
8. Barrett, H., Myers, K.: Foundations of Image Science. Wiley Series in Pure and Applied Optics. Wiley, Hoboken (2004)
9. Beard, P.C., Laufer, J.G., Cox, B., Arridge, S.R.: Quantitative photoacoustic imaging: measurement of absolute chromophore concentrations for physiological and molecular imaging. In: Wang, L.V. (ed.) Photoacoustic Imaging and Spectroscopy. CRC, Boca Raton (2009)
10. Bertero, M., Boccacci, P.: Inverse Problems in Imaging. Institute of Physics Publishing, Bristol (1998)
11. Cheong, W., Prahl, S., Welch, A.: A review of the optical properties of biological tissues. IEEE J. Quantum Electron. **26**, 2166–2185 (1990)
12. Cox, B.T., Arridge, S.R., Kstli, K.P., Beard, P.C.: Two-dimensional quantitative photoacoustic image reconstruction of absorption distributions in scattering media by use of a simple iterative method. Appl. Opt. **45**, 1866–1875 (2006)
13. Devaney, A.J.: The inverse problem for random sources. J. Math. Phys. **20**, 1687–1691 (1979)
14. Devaney, A.J.: Inverse source and scattering problems in ultrasonics. IEEE Trans. Sonics Ultrason. **30**, 355–364 (1983)
15. Diebold, G.J.: Photoacoustic monopole radiation: waves from objects with symmetry in one, two, and three dimension. In: Wang, L.V. (ed.) Photoacoustic Imaging and Spectroscopy. CRC, Boca Raton (2009)
16. Diebold, G.J., Westervelt, P.J.: The photoacoustic effect generated by a spherical droplet in a fluid. J. Acoust. Soc. Am. **84**(6), 2245–2251 (1988)
17. Diebold, G.J., Sun, T., Khan, M.I.: Photoacoustic monopole radiation in one, two, and three dimensions. Phys. Rev. Lett. **67**(24), 3384–3387 (1991)
18. Ephrat, P., Keenliside, L., Seabrook, A., Prato, F.S., Carson, J.J.L.: Three-dimensional photoacoustic imaging by sparse-array detection and iterative image reconstruction. J. Biomed. Opt. **13**(5), 054052 (2008)
19. Esenaliev, R.O., Karabutov, A.A., Oraevsky, A.A.: Sensitivity of laser opto-acoustic imaging in detection of small deeply embedded tumors. IEEE J. Sel. Top. Quantum Electron. **5**, 981–988 (1999)
20. Fessler, J.A.: Penalized weighted least-squares reconstruction for positron emission tomography. IEEE Trans. Med. Imaging **13**, 290–300 (1994)
21. Fessler, J.A., Booth, S.D.: Conjugate-gradient preconditioning methods for shiftvariant PET image reconstruction. IEEE Trans. Image Process. **8**(5), 688–699 (1999)
22. Finch, D., Patch, S., Rakesh, M.: Determining a function from its mean values over a family of spheres. SIAM J. Math. Anal. **35**, 1213–1240 (2004)
23. Finch, D., Haltmeier, M., Rakesh, M.: Inversion of spherical means and the wave equation in even dimensions. SIAM J. Appl. Math. **68**(2), 392–412 (2007)
24. Haltmeier, M., Scherzer, O., Burgholzer, P., Paltauf, G.: Thermoacoustic computed tomography with large planar receivers. Inverse Probl. **20**(5), 1663–1673 (2004)
25. Jin, X., Wang, L.V.: Thermoacoustic tomography with correction for acoustic speed variations. Phys. Med. Biol. **51**(24), 6437–6448 (2006)
26. Joines, W., Jirtle, R., Rafal, M., Schaeffer, D.: Microwave power absorption differences between normal and malignant tissue. Radiat. Oncol. Biol. Phys. **6**, 681–687 (1980)
27. Khokhlova, T.D., Pelivanov, I.M., Kozhushko, V.V., Zharinov, A.N., Solomatin, V.S., Karabutov, A.A.: Optoacoustic imaging of absorbing objects in a turbid medium: ultimate sensitivity and application to breast cancer diagnostics. Appl. Opt. **46**(2), 262–272 (2007)
28. Köstli, K.P., Beard, P.C.: Two-dimensional photoacoustic imaging by use of fouriertransform image reconstruction and a detector with an anisotropic response. Appl. Opt. **42**(10), 1899–1908 (2003)
29. Köstli, K.P., Frenz, M., Bebie, H., Weber, H.P.: Temporal backward projection of optoacoustic pressure transients using Fourier transform methods. Phys. Med. Biol. **46**(7), 1863–1872 (2001)

30. Kruger, R.A., Liu, P., Fang, R., Appledorn, C.: Photoacoustic ultrasound (PAUS) reconstruction tomography. Med. Phys. **22**, 1605–1609 (1995)
31. Kruger, R., Reinecke, D., Kruger, G.: Thermoacoustic computed tomography-technical considerations. Med. Phys. **26**, 1832–1837 (1999)
32. Kruger, R.A., Kiser, W.L., Reinecke, D.R., Kruger, G.A., Miller, K.D.: Thermoacoustic optical molecular imaging of small animals. Mol. Imaging **2**, 113–123 (2003)
33. Ku, G., Fornage, B.D., Jin, X., Xu, M., Hunt, K.K., Wang, L.V.: Thermoacoustic and photoacoustic tomography of thick biological tissues toward breast imaging. Technol. Cancer Res. Treat. **4**, 559–566 (2005)
34. Kuchment, P., Kunyansky, L.: Mathematics of thermoacoustic tomography. Eur. J. Appl. Math. **19**, 191–224 (2008)
35. Kunyansky, L.A.: Explicit inversion formulae for the spherical mean radon transform. Inverse Probl. **23**, 373–383 (2007)
36. Langenberg, K.J.: Basic Methods of Tomography and Inverse Problems. Adam Hilger, Philadelphia (1987)
37. La Riviere, P.J., Zhang, J., Anastasio, M.A.: Image reconstruction in optoacoustic tomography for dispersive acoustic media. Opt. Lett. **31**, 781–783 (2006)
38. Lewitt, R.M.: Alternatives to voxels for image representation in iterative reconstruction algorithms. Phys. Med. Biol. **37**(3), 705–716 (1992)
39. Li, C., Wang, L.V.: Photoacoustic tomography and sensing in biomedicine. Phys. Med. Biol. **54**(19), R59–R97 (2009)
40. Li, C., Pramanik, M., Ku, G., Wang, L.V.: Image distortion in thermoacoustic tomography caused by microwave diffraction. Phys. Rev. E Stat. Nonlinear Soft Matter Phys. **77**(3), 031923 (2008)
41. Maslov, K., Wang, L.V.: Photoacoustic imaging of biological tissue with intensitymodulated continuous-wave laser. J. Biomed. Opt. **13**(2), 024006 (2008)
42. Modgil, D., Anastasio, M.A., Wang, K., LaRivière, P.J.: Image reconstruction in photoacoustic tomography with variable speed of sound using a higher order geometrical acoustics approximation. In: SPIE, vol. 7177, p. 71771A (2009)
43. Norton, S., Linzer, M.: Ultrasonic reflectivity imaging in three dimensions: exact inverse scattering solutions for plane, cylindrical, and spherical apertures. IEEE Trans. Biomed. Eng. **28**, 202–220 (1981)
44. Oraevsky, A.A., Karabutov, A.A.: Ultimate sensitivity of time-resolved optoacoustic detection. In: SPIE, vol. 3916, pp 228–239 (2000)
45. Oraevsky, A.A., Karabutov, A.A.: Optoacoustic tomography. In: Vo-Dinh, T. (ed.) Biomedical Photonics Handbook. CRC, Boca Raton (2003)
46. Oraevsky, A.A., Jacques, S.L., Tittel, F.K.: Measurement of tissue optical properties by time-resolved detection of laser-induced transient stress. Appl. Opt. **36**, 402–415 (1997)
47. Paltauf, G., Schmidt-Kloiber, H., Guss, H.: Light distribution measurements in absorbing materials by optical detection of laser-induced stress waves. Appl. Phys. Lett. **69**(11), 1526–1528 (1996)
48. Paltauf, G., Viator, J., Prahl, S., Jacques, S.: Iterative reconstruction algorithm for optoacoustic imaging. J. Acoust. Soc. Am. **112**, 1536–1544 (2002)
49. Paltauf, G., Nuster, R., Burgholzer, P.: Characterization of integrating ultrasound detectors for photoacoustic tomography. J. Appl. Phys. **105**(10), 102026 (2009)
50. Pan, X., Zou, Y., Anastasio, M.A.: Data redundany and reduced-scan reconstruction in reflectivity tomography. IEEE Trans. Image Process. **12**, 784–795 (2003)
51. Patch, S.K.: Thermoacoustic tomography–consistency conditions and the partial scan problem. Phys. Med. Biol. **49**(11), 2305–2315 (2004)
52. Provost, J., Lesage, F.: The application of compressed sensing for photo-acoustic tomography. IEEE Trans. Med. Imaging **28**, 585–594 (2009)
53. Sushilov, N.V., Cobbold, S.C.: Frequency-domain wave equation and its timedomain solutions in attenuating media. J. Acoust. Soc. Am. **115**(4), 1431–1436 (2004)

54. Tam, A.C.: Application of photo-acoustic sensing techniques. Rev. Mod. Phys. **58**, 381–431 (1986)
55. Wang, L.V. (ed.): Photoacoustic Imaging and Spectroscopy. CRC, Boca Raton (2009)
56. Wang, L.V., Wu, H.-I.: Biomedical Optics, Principles and Imaging. Wiley, Hoboken (2007)
57. Wang, L.V., Zhao, X.M., Sun, H.T., Ku, G.: Microwave-induced acoustic imaging of biological tissues. Rev. Sci. Instrum. **70**, 3744–3748 (1999)
58. Wang, Y., Xie, X., Wang, X., Ku, G., Gill, K.L., ONeal, D.P., Stoica, G., Wang, L.V.: Photoacoustic tomography of a nanoshell contrast agent in the in vivo rat brain. Nano Lett. **4**, 1689–1692 (2004)
59. Wang, X., Xie, X., Ku, G., Wang, L.V., Stoica, G.: Noninvasive imaging of hemoglobin concentration and oxygenation in the rat brain using high-resolution photoacoustic tomography. J. Biomed. Opt. **11**(2), 024015 (2006)
60. Wernick, M.N., Aarsvold, J.N.: Emission Tomography, the Fundamentals of PET and SPECT. Elsevier, San Diego (2004)
61. Xu, M., Wang, L.V.: Time-domain reconstruction for thermoacoustic tomography in a spherical geometry. IEEE Trans. Med. Imaging **21**, 814–822 (2002)
62. Xu, M., Wang, L.V.: Analytic explanation of spatial resolution related to bandwidth and detector aperture size in thermoacoustic or photoacoustic reconstruction. Phys. Rev. E **67**, 056605 (2003)
63. Xu, Y., Wang, L.V.: Effects of acoustic heterogeneity in breast thermoacoustic tomography. IEEE Trans. Ultrason. Ferroelectr. Freq. Control **50**, 1134–1146 (2003)
64. Xu, M., Wang, L.: Universal back-projection algorithm for photoacoustic computed tomography. Phys. Rev. E **71**, 016706 (2005)
65. Xu, M., Wang, L.V.: Biomedical photoacoustics. Rev. Sci. Instrum. **77**, 041101 (2006)
66. Xu, Y., Feng, D., Wang, L.V.: Exact frequency-domain reconstruction for thermoacoustic tomography i: planar geometry. IEEE Trans. Med. Imaging **21**, 823–828 (2002)
67. Xu, Y., Xu, M., Wang, L.V.: Exact frequency-domain reconstruction for thermoacoustic tomography-ii: cylindrical geometry. IEEE Trans. Med. Imaging **21**, 829–833 (2002)
68. Yuan, Z., Jiang, H.: Quantitative photoacoustic tomography: recovery of optical absorption coefficient maps of heterogeneous media. Appl. Phys. Lett. **88**(23), 231101 (2006)
69. Zhang, J., Anastasio, M.A., Pan, X., Wang, L.V.: Weighted expectation maximization reconstruction algorithms for thermoacoustic tomography. IEEE Trans. Med. Imaging **24**, 817–820 (2005)
70. Zou, Y., Pan, X., Anastasio, M.A.: Data truncation and the exterior reconstruction problem in reflection-mode tomography. In: IEEE Nuclear Science Symposium Conference Record, Norfolk, vol. 2, pp. 726–730. IEEE, (2002)

Mathematics of Photoacoustic and Thermoacoustic Tomography

Peter Kuchment and Leonid Kunyansky

Contents

1 Introduction 1118
2 Mathematical Models of TAT 1119
Point Detectors and the Wave Equation Model 1119
Acoustically Homogeneous Media and Spherical Means 1121
Main Mathematical Problems Arising in TAT 1122
Variations on the Theme: Planar, Linear, and Circular Integrating Detectors 1123
3 Mathematical Analysis of the Problem 1126
Uniqueness of Reconstruction 1126
Stability 1133
Incomplete Data 1135
Discussion of the Visibility Condition 1139
Range Conditions 1140
Reconstruction of the Speed of Sound 1144
4 Reconstruction Formulas, Numerical Methods, and Case Examples 1145
Full Data (Closed Acquisition Surfaces) 1146
Partial (Incomplete) Data 1157
5 Final Remarks and Open Problems 1161
Cross-References 1163
References 1163

P. Kuchment (✉)
Mathematics Department, Texas A & M University, College Station, TX, USA
e-mail: kuchment@math.tamu.edu

L. Kunyansky
Department of Mathematics, University of Arizona, Tucson, AZ, USA
e-mail: leonk@math.arizona.edu

© Springer Science+Business Media New York 2015

O. Scherzer (ed.), *Handbook of Mathematical Methods in Imaging*,
DOI 10.1007/978-1-4939-0790-8_51

Abstract

The chapter surveys the mathematical models, problems, and algorithms of the thermoacoustic tomography (TAT) and photoacoustic tomography (PAT). TAT and PAT represent probably the most developed of the several novel "hybrid" methods of medical imaging. These new modalities combine different physical types of waves (electromagnetic and acoustic in case of TAT and PAT) in such a way that the resolution and contrast of the resulting method are much higher than those achievable using only acoustic or electromagnetic measurements.

1 Introduction

We provide here just a very brief description of the thermoacoustic tomography/photoacoustic tomography (TAT/PAT) procedure, since the relevant physics and biology details can be found in another chapter [94] in this volume, as well as in the surveys and books [93, 95]. In TAT (PAT), a short pulse of radio-frequency EM wave (correspondingly, laser beam) irradiates a biological object (e.g., in the most common application, human breast), thus causing small levels of heating. The resulting thermoelastic expansion generates a pressure wave that starts propagating through the object. The absorbed EM energy and the initial pressure it creates are much higher in the cancerous cells than in healthy tissues (see the discussion of this effect in [93–95]). Thus, if one could reconstruct the initial pressure $f(x)$, the resulting TAT tomogram would contain highly useful diagnostic information. The data for such a reconstruction are obtained by measuring time-dependent pressure $p(x,t)$ using acoustic transducers located on a surface S (we will call it the *observation* or *acquisition surface*) completely or partially surrounding the body (see Fig. 1). Thus, although the initial irradiation is electromagnetic, the actual reconstruction is based on acoustic measurements. As a result, the high contrast is produced due to a much higher absorption of EM energy by cancerous cells (ultrasound alone would not produce good contrast in this case), while the good (submillimeter) resolution is achieved by using ultrasound measurements (the radio-frequency EM waves are too long for high-resolution imaging). Thus, TAT, by using two types of waves, combines their advantages, while eliminating their individual deficiencies.

The physical principle upon which TAT/PAT is based was discovered by Alexander Graham Bell in 1880 [19], and its application for imaging of biological tissues was suggested a century later [21]. It began to be developed as a viable medical imaging technique in the middle of the 1990s [53, 69].

Some of the mathematical foundations of this imaging modality were originally developed starting in the 1990s for the purposes of the approximation theory, integral geometry, and sonar and radar (see [4, 7, 38, 54, 60] for references and extensive reviews of the resulting developments). Physical, biological, and mathematical aspects of TAT/PAT have been recently reviewed in [4, 38, 39, 54, 70, 89, 92, 93, 95].

TAT/PAT is just one, probably the most advanced at the moment, example of the several recently introduced hybrid imaging methods, which combine different

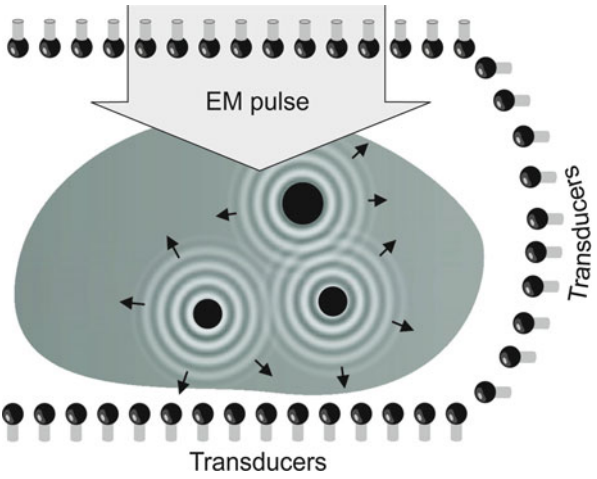


Fig. 1 Thermoacoustic tomography/photoacoustic tomography (TAT/PAT) procedure with a partially surrounding acquisition surface

types of radiation to yield high quality of imaging unobtainable by single-radiation modalities (see [10, 11, 40, 55, 95] for other examples).

2 Mathematical Models of TAT

In this section, we describe the commonly accepted mathematical model of the TAT procedure and the main mathematical problems that need to be addressed. Since for all our purposes PAT results in the same mathematical model (although the biological features that TAT and PAT detect are different; see details in Ref. [13]), we will refer to TAT only.

Point Detectors and the Wave Equation Model

We will mainly assume that point-like omnidirectional ultrasound transducers, located throughout an observation (acquisition) surface S, are used to detect the values of the pressure $p(y,t)$, where $y \in S$ is a detector location and $t \geq 0$ is the time of the observation. We also denote by $c(x)$ the speed of sound at a location x. Then, it has been argued that the following model describes correctly the propagating pressure wave $p(x,t)$ generated during the TAT procedure (e.g., [13, 31, 88, 94, 97]):

$$\begin{cases} p_{tt} = c^2(x)\Delta_x p, \quad t \geq 0, x \in \mathbb{R}^3 \\ p(x,0) = f(x), \, p_t(x,0) = 0. \end{cases} \tag{1}$$

Here $f(x)$ is the initial value of the acoustic pressure, which one needs to find in order to create the TAT image. In the case of a closed acquisition surface S, we will

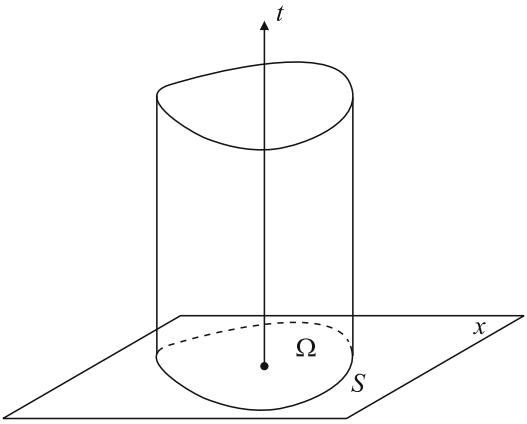


Fig. 2 The observation surface S and the domain Ω containing the object to be imaged

denote by Ω the interior domain it bounds. Notice that in TAT the function $f(x)$ is naturally supported inside Ω. We will see that this assumption about the support of f sometimes becomes crucial for the feasibility of reconstruction, although some issues can be resolved even if f has nonzero parts outside the acquisition surface.

The data obtained by the point detectors located on a surface S are represented by the function

$$g(y,t) := p(y,t) \quad \text{for } y \in S, t \geq 0. \tag{2}$$

Figure 2 illustrates the space-time geometry of (1).

We will incorporate the measured data g into the system (1), rewriting it as follows:

$$\begin{cases} p_{tt} = c^2(x)\Delta_x p, \quad t \geq 0, x \in \mathbb{R}^3 \\ p(x,0) = f(x), \, p_t(x,0) = 0 \\ p|_S = g(y,t), \quad (y,t) \in S \times \mathbb{R}^+. \end{cases} \tag{3}$$

Thus, the goal in TAT/PAT is to find, using the data $g(y,t)$ measured by transducers, the initial value $f(x)$ at $t = 0$ of the solution $p(x,t)$ of (3).

We will use the following notation:

Definition 1. We will denote by $\mathcal{W}$ the forward operator

$$\mathcal{W} : f(x) \mapsto g(y,t), \tag{4}$$

where f and g are described in (3).

Remark 1.

- The reader should notice that if a different type of detector is used, the system (1) will still hold, while the measured data will be represented differently from (2) (see section "Variations on the Theme: Planar, Linear, and Circular Integrating Detectors"). This will correspondingly influence the reconstruction procedures.
- We can consider the same problem in the space $\mathbb{R}^n$ of any dimension, not just in $3D$. This is not merely a mathematical abstraction. Indeed, in the case of the so-called integrating line detectors (section "Variations on the Theme: Planar, Linear, and Circular Integrating Detectors"), one deals with the $2D$ situation.

Acoustically Homogeneous Media and Spherical Means

If the medium being imaged is acoustically homogeneous (i.e., $c(x)$ equals to a constant, which we will assume to be equal to 1 in appropriate units), as it is approximately the case in breast imaging, one deals with the constant coefficient wave equation problem

$$\begin{cases} p_{tt} = \Delta_x p, \quad t \geq 0, x \in \mathbb{R}^3 \\ p(x,0) = f(x), p_t(x,0) = 0 \\ p|_S = g(y,t), \quad (y,t) \in S \times \mathbb{R}^+. \end{cases} \tag{5}$$

In this case, the well-known Poisson–Kirchhoff formulas [27, Chap. VI, Sect. 13.2, Formula (15)] for the solution of the wave equation give in $3D$

$$p(x,t) = a\frac{\partial}{\partial t}\left(t(Rf)(x,t)\right), \tag{6}$$

where

$$(Rf)(x,r) := \frac{1}{4\pi}\int\limits_{|y|=1} f(x+ry)dA(y) \tag{7}$$

is the spherical mean operator applied to the function $f(x)$, dA is the standard area element on the unit sphere in $\mathbb{R}^3$, and a is a constant. (Versions in all dimensions are known; see (16) and (15).) One can derive from here that knowledge of the function $g(x,t)$ for $x \in S$ and all $t \geq 0$ is equivalent to knowing the spherical mean $Rf(x,t)$ of the function f for any points $x \in S$ and any $t \geq 0$. One thus needs to study the spherical mean operator $R : f \to Rf$, or, more precisely, its restriction to the points $x \in S$ only, which we will denote by $\mathcal{M}$:

$$\mathcal{M}f(x,t) := \frac{1}{4\pi}\int\limits_{|y|=1} f(x+ty)dA(y), \quad x \in S, t \geq 0. \tag{8}$$

Due to the connection between the spherical mean operator and the wave equation, one can choose to work with the former, and in fact many works on TAT do so. The spherical mean operator $\mathcal{M}$ resembles the classical Radon transform, the common tool of computed tomography [63], which integrates functions over planes rather than spheres. This analogy with Radon transform, although often purely ideological, rather than technical, provides important intuition and frequently points in reasonable directions of study. However, when the medium cannot be assumed to be acoustically homogeneous, and thus $c(x)$ is not constant, the relation between TAT and integral geometric transforms, such as the Radon and spherical mean transforms to a large extent breaks down, and thus one has to work with the wave equation directly.

In what follows, we will address both models of TAT (the PDE model and the integral geometry model) and thus will deal with both forward operators $\mathcal{W}$ and $\mathcal{M}$.

Main Mathematical Problems Arising in TAT

We now formulate a list of problems related to TAT, which will be addressed in detail in the rest of the article. (This list is more or less standard for a tomographic imaging method.)

Sufficiency of the data: The first natural question to ask is as follows: Is the data collected on the observation surface S sufficient for the unique reconstruction of the initial pressure $f(x)$ in (3)? In other words, is the kernel of the forward operator $\mathcal{W}$ zero? Or, to put it differently, for which sets $S \in \mathbb{R}^3$ the data collected by transducers placed along S determines f uniquely? Yet another interpretation of this question is through observability of solutions of the wave equation on the set S: does observation on S of a solution of the problem (1) determine the solution uniquely?

When the speed of sound is constant, and thus the spherical mean model applies, the equivalent question is whether the operator $\mathcal{M}$ has zero kernel on an appropriate class of functions (say, continuous functions with compact support). As it is explained in [7], the choice of precise conditions on the local function class, such as continuity, is of no importance for the answer to the uniqueness question, while behavior at infinity (e.g., compactness of support) is. So, without loss of generality, when discussing uniqueness, one can assume $f(x)$ in (3) to be infinitely differentiable.

Inversion formulas and algorithms: Since a practitioner needs to see the actual tomogram, rather than just know its existence, the next natural question arises: If uniqueness the data collected on S is established, what are the actual inversion formulas or algorithms? Here again one can work with smooth functions, in the end extending the formulas by continuity to a wider class.

Stability of reconstruction: If we can invert the transform and reconstruct f from the data g, how stable is the inversion? The measured data are unavoidably

corrupted by errors, and stability means that small errors in the data lead to only small errors in the reconstructed tomogram.

Incomplete data problems: What happens if the data is "incomplete," for instance, if one can only partially surround the object by transducers? Does this lead to any specific deterioration in the tomogram and if yes, to what kind of deterioration?

Range descriptions: The next question is known to be important for the analysis of tomographic problems: What is the range of the forward operator $\mathcal{W} : f \mapsto g$ that maps the unknown function f to the measured data g? In other words, what is the space of all possible "ideal" data $g(t, y)$ collected on the surface S? In the constant speed of sound case, this is equivalent to the question of describing the range of the spherical mean operator $\mathcal{M}$ in appropriate function spaces. Such ranges often have infinite co-dimensions, and the importance of knowing the range of Radon type transforms for analyzing problems of tomography is well known. For instance, such information is used to improve inversion algorithms, complete incomplete data, and discover and compensate for certain data errors (e.g., [41, 45, 63, 68, 70] and references therein). In TAT, range descriptions are also closely connected with the speed of sound determination problem listed next (see section "Reconstruction of the Speed of Sound" for a discussion of this connection).

Speed of sound reconstruction: As the reader can expect, reconstruction procedures require the knowledge of the speed of sound $c(x)$. Thus, the problem arises of the recovery of $c(x)$ either from an additional scan or (preferably) from the same TAT data.

Variations on the Theme: Planar, Linear, and Circular Integrating Detectors

In the most basic and well-studied version of TAT described above, one utilizes point-like broadband transducers to measure the acoustic wave on a surface surrounding the object of interest. The corresponding mathematical model is described by the system (3). In practice, the transducers cannot be made small enough, since smaller detectors yield weaker signals resulting in low signal-to-noise ratios. Smaller transducers are also more difficult to manufacture.

Since finite size of the transducers limits the resolution of the reconstructed images, researchers have been trying to design alternative acquisition schemes using receivers that are very thin but long or wide. Such are $2D$ planar detectors [24, 43] and $1D$ linear and circular [23, 42, 73, 103] detectors.

We will assume throughout this section that the speed of sound $c(x)$ is constant and equal to 1.

Planar detectors are made from a thin piezoelectric polymer film glued onto a flat substrate (see, e.g., [75]). Let us assume that the object is contained within the sphere of radius R. If the diameter of the planar detector is sufficiently large

(see [75] for details), it can be assumed to be infinite. The mathematical model of such an acquisition technique is no longer described by (3). Let us define the detector plane $\Pi(s,\omega)$ by equation $x\cdot\omega = s$, where ω is the unit normal to the plane and s is the (signed) distance from the origin to the plane. Then, while the propagation of acoustic waves is still modeled by (1), the measured data $g_{\text{planar}}(s,t,\omega)$ (up to a constant factor which we will, for simplicity, assume to be equal to 1) can be represented by the following integral:

$$g_{\text{planar}}(s,\omega,t) = \int\limits_{\Pi(s,\omega)} p(x,t)dA(x)$$

where $dA(x)$ is the surface measure on the plane. Obviously,

$$g_{\text{planar}}(s,\omega,0) = \int\limits_{\Pi(s,\omega)} p(x,0)dA(x) = \int\limits_{\Pi(s,\omega)} f(x)dA(x) \equiv F(s,\omega),$$

i.e., the value of g at $t = 0$ coincides with the integral $F(s,\omega)$ of the initial pressure $f(x)$ over the plane $\Pi(s,\omega)$ orthogonal to ω.

One can show [24, 43] that for a fixed ω, function $g_{\text{planar}}(s,\omega,t)$ is the solution to $1D$ wave equation

$$\frac{\partial^2 g}{\partial s^2} = \frac{\partial^2 g}{\partial t^2},$$

and thus

$$\begin{aligned} g_{\text{planar}}(s,\omega,t) &= \frac{1}{2}\left[g_{\text{planar}}(s,\omega,s-t) + g_{\text{planar}}(s,\omega,s+t)\right] \\ &= \frac{1}{2}\left[F(s+t,\omega) + F(s-t,\omega)\right]. \end{aligned}$$

Since the detector can only be placed outside the object, i.e., $s \geq R$, the term $F(s+t,\omega)$ vanishes, and one obtains

$$g_{\text{planar}}(s,\omega,t) = F(s-t,\omega).$$

In other words, by measuring $g_{\text{planar}}(s,\omega,t)$, one can obtain values of the planar integrals of $f(x)$. If, as proposed in [24, 43], one conducts measurements for all planes tangent to the upper half sphere of radius R (i.e., $s = R, \omega \in S^2_+$), then the resulting data yield all values of the standard Radon transform of $f(x)$. Now the reconstruction can be carried out using one of the many known inversion algorithms for the latter transform [63].

Linear detectors are based on optical detection of acoustic signal. Some of the proposed optical detection schemes utilize as the sensitive element a thin straight optical fiber in combination with Fabry–Perot interferometer [23, 42]. Changes of

acoustic pressure on the fiber change (proportionally) its length; this elongation, in turn, is detected by the interferometer. A similar idea is used in [73]; in this work the role of a sensitive element is played by a laser beam passing through the water in which the object of interest is submerged, and thus the measurement does not perturb the acoustic wave. In both cases, the length of the sensitive element exceeds the size of the object, while the diameter of the fiber (or of the laser beam) can be made extremely small (see [75] for a detailed discussion), which removes restrictions on resolution one can achieve in the images.

Let us assume that the fiber (or laser beam) is aligned along the line $l(s_1, s_2, \omega_1, \omega_2) = \{x | x = s_1\omega_1 + s_2\omega_2 + s\omega\}$, where vectors ω_1, ω_2, and ω form an orthonormal basis in $\mathbb{R}^3$. Then the measured quantities $g_{\text{linear}}(s_1, s_2, \omega_1, \omega_2, t)$ are equal (up to a constant factor which, we will assume, equals to 1) to the following line integral:

$$g_{\text{linear}}(s_1, s_2, \omega_1, \omega_2, t) = \int_{\mathbb{R}^1} p(s_1\omega_1 + s_2\omega_2 + s\omega, t)ds.$$

Similar to the case of planar detection, one can show [23,42,73] that for fixed vectors ω_1, ω_2 the measurements $g_{\text{linear}}(s_1, s_2, \omega_1, \omega_2, t)$ satisfy the $2D$ wave equation

$$\frac{\partial^2 g}{\partial s_1^2} + \frac{\partial^2 g}{\partial s_2^2} = \frac{\partial^2 g}{\partial t^2}.$$

The initial values $g_{\text{linear}}(s_1, s_2, \omega_1, \omega_2, 0)$ coincide with the line integrals of $f(x)$ along lines $l(s_1, s_2, \omega_1, \omega_2)$. Suppose one makes measurements for all values of $s_1(\tau)$, $s_2(\tau)$ corresponding to a curve $\gamma = \{x | x = s_1(\tau)\omega_1 + s_2(\tau)\omega_2, \tau_0 \le \tau \le \tau_1\}$ lying in the plane spanned by ω_1, ω_2. Then one can try to reconstruct the initial value of g from the values of g on γ. This problem is a $2D$ version of (3), and thus the known algorithms (see Sect. 4) are applicable.

In order to complete the reconstruction from data obtained using line detectors, the measurements should be repeated with different directions of ω. For each value of ω, the $2D$ problem is solved; the solutions of these problems yield values of line integrals of $f(x)$. If this is done for all values of ω lying on a half circle, the set of the recovered line integrals of $f(x)$ is sufficient for reconstructing this function. Such a reconstruction represents the inversion of the well known in tomography X-ray transform. The corresponding theory and algorithms can be found, for instance, in [63].

Finally, the use of circular integrating detectors was considered in [103]. Such a detector can be made out of optical fiber combined with an interferometer. In [103], a closed-form solution of the corresponding inverse problem is found. However, this approach is very new, and neither numerical examples nor reconstructions from real data have been obtained yet.

3 Mathematical Analysis of the Problem

In this section, we will address most of the issues described in section "Main Mathematical Problems Arising in TAT," except the reconstruction algorithms, which will be discussed in Sect. 4.

Uniqueness of Reconstruction

The problem discussed here is the most basic one for tomography: Given an acquisition surface S along which we distribute detectors, is the data $g(y,t)$ for $y \in S, t \geq 0$ (see (3)) sufficient for a unique reconstruction of the tomogram f? A simple counting of variables shows that S should be a hypersurface in the ambient space (i.e., a surface in $\mathbb{R}^3$ or a curve in $\mathbb{R}^2$). As we will see below, although there are some simple counterexamples and remaining open problems, for all practical purposes, the uniqueness problem is positively resolved, and most surfaces S do provide uniqueness. We address this issue for acoustically homogeneous media first and then switch to the variable speed case.

Before doing so, however, we would like to dispel a concern that arises when one looks at the problem of recovering f from g in (3). Namely, an impression might arise that we consider an initial-boundary value (IBV) problem for the wave equation in the cylinder $\Omega \times \mathbb{R}^+$, and the goal is to recover the initial data f from the known boundary data g. This is clearly impossible, since according to standard PDE theorems (e.g., [27]), one can solve this IBV problem for **arbitrary** choice of the initial data f and boundary data g (as long as they satisfy simple compatibility conditions, which are fulfilled, for instance, if f vanishes near S and g vanishes for small t, which is the case in TAT). This means that apparently g contains essentially no information about f at all. This argument, however, is flawed, since the wave equation in (3) holds in the whole space, not just in Ω. In other words, S is not a boundary, but rather an observation surface. In particular, considering the wave equation in the exterior of S, one can derive that if f is supported inside Ω, the boundary values g of the solution p of (3) also determine the normal derivative of p at S for all positive times. Thus, we in fact have (at least theoretically) the full Cauchy data of the solution p on S, which should be sufficient for reconstruction. Another way of addressing this issue is to notice that if the speed of sound is constant, or at least non-trapping (see the definition below in section "Acoustically Inhomogeneous Media"), the energy of the solution in any bounded domain (in particular, in Ω) must decay in time. The decay when $t \to \infty$ together with the boundary data g guarantees the uniqueness of solution and thus uniqueness of recovery f.

These arguments, as the reader will see, play a role in understanding reconstruction procedures.

Acoustically Homogeneous Media

We assume here the sound speed $c(x)$ to be constant (in appropriate units, one can choose it to be equal to 1, which we will do to simplify considerations).

In order to state the first important result on uniqueness, let us recall the system (5), allowing an arbitrary dimension n of the space:

$$\begin{cases} p_{tt} = \Delta_x p, \quad t \geq 0, x \in \mathbb{R}^n \\ p(x,0) = f(x), \, p_t(x,0) = 0 \\ p|_S = g(y,t), \quad (y,t) \in S \times \mathbb{R}^+. \end{cases} \tag{9}$$

We introduce the following useful definition:

Definition 2. A set S is said to be a uniqueness set if when used as the acquisition surface, it provides sufficient data for unique reconstruction of the compactly supported tomogram f (i.e., the observed data g in (9) determines uniquely function f). Otherwise, it is called a nonuniqueness set.

In other words, S is a uniqueness set if the forward operator $\mathcal{W}$ (or, equivalently, $\mathcal{M}$) has zero kernel.

We have not indicated above the smoothness class of $f(x)$. However, it is not hard to show (e.g., [7]) that the uniqueness does not depend on the smoothness of f; for simplicity, the reader can assume that f is infinitely differentiable. On the other hand, compactness of support is important in what follows.

We will start with a very general statement about the acquisition (observation) sets S that provide insufficient information for unique reconstruction of f (see [7] for the proof and references):

Theorem 1. *If S is a nonuniqueness set, then there exists a nonzero harmonic polynomial Q, which vanishes on S.*

This theorem implies, in particular, that all "bad" (nonuniqueness) observation sets are algebraic, i.e., they have a polynomial vanishing on them. Turning this statement around, we conclude that any set S that is a uniqueness set for harmonic polynomials is sufficient for unique TAT reconstruction (although, as we will see in section "Incomplete Data," this does not mean practicality of the reconstruction).

The proof of Theorem 1, which the reader can find in [7, 54], is not hard and in fact is enlightening, but providing it would lead us too far from the topic of this survey.

We will consider first the case of closed acquisition surfaces, i.e., the ones that completely surround the object to be imaged. We will address the general situation afterward.

Closed Acquisition Surfaces S

Theorem 2 ([7]). *If the acquisition surface S is the boundary of bounded domain Ω (i.e., a closed surface), then it is a uniqueness set. Thus, the observed data g in (9) determines uniquely the sought function $f \in L^2_{comp}(\mathbb{R}^n)$. (The statement holds, even though f is not required to be supported inside S.)*

Proof. Indeed, since there are no nonzero harmonic functions vanishing on a closed surface S, Theorem 1 implies Theorem 2. ■

There is, however, another more intuitive explanation of why Theorem 2 holds true (although it requires somewhat stronger assumptions or a more delicate proof than the one indicated below). Namely, since the solution p of (9) has compactly supported initial data, its energy is decaying inside any bounded domain, in particular inside Ω (see section "Acoustically Inhomogeneous Media" and [32,47] and references therein about local energy decay). On the other hand, if there is nonuniqueness, there exists a nonzero f such that $g(y,t) = 0$ for all $y \in S$ and t. This means that we can add homogeneous Dirichlet boundary conditions $p\,|_S = 0$ to (9). But then the standard PDE theorems [27] imply that the energy stays constant in Ω. Combination of the two conclusions means that p is zero in Ω for all times t. It is well known [27] that such a solution of the wave equation must be identically zero everywhere, and thus $f = 0$.

This energy decay consideration can be extended to some classes of non-compactly supported functions f of the L^p classes, leading to the following result of [1]:

Theorem 3 ([1]). *Let S be the boundary of a bounded domain in $\mathbb{R}^n$ and $f \in L^p(\mathbb{R}^n)$. Then:*

1. *If $p \le \frac{2n}{n-1}$ and the spherical mean of f over almost every sphere centered on S is equal to zero, then $f = 0$.*
2. *The previous statement fails when $p > \frac{2n}{n-1}$ and S is a sphere.*

In other words, a closed surface S is a uniqueness set for functions $f \in L^p(\mathbb{R}^n)$ when $p \le \frac{2n}{n-1}$ and might fail to be such when $p > \frac{2n}{n-1}$.

This result shows that the assumption, if not necessarily of compactness of support of f, but at least of a sufficiently fast decay of f at infinity, is important for the uniqueness to hold.

General Acquisition Sets S

Theorems 1 and 2 imply the following useful statement:

Theorem 4. *If a set S is not algebraic, or if it contains an open part of a closed analytic surface Γ, then it is a uniqueness set.*

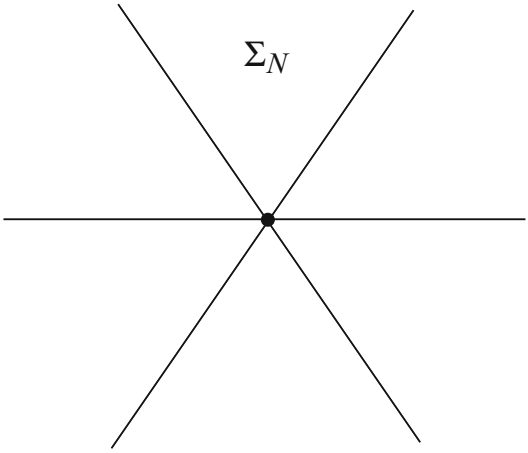


Fig. 3 Coxeter cross of N lines

Indeed, the first claim follows immediately from Theorem 1. The second one works out as follows: if an open subset of an analytic surface Γ is a nonuniqueness set, then by an analytic continuation type argument (see [7]), one can show that the whole Γ is such a set. However, this is impossible, due to Theorem 2.

There are simple examples of nonuniqueness surfaces. Indeed, if S is a plane in $3D$ (or a line in $2D$ or a hyperplane in dimension n) and $f(x)$ in (3) is odd with respect to S, then clearly the whole solution of (3) has the same parity and thus vanishes on S for all times t. This means that, if one places transducers on a planar S, they might register zero signals at all times, while the function f to be reconstructed is not zero. Thus, there is no uniqueness of reconstruction when S is a plane. On the other hand (see [27, 51]), if f is supported completely on one side of the plane S (the standard situation in TAT), it is uniquely recoverable from its spherical means centered on S and thus from the observed data g.

The question arises on what are other "bad" (nonuniqueness) acquisition surfaces than planes. This issue has been resolved in $2D$ only. Namely, consider a set of N lines on the plane intersecting at a point and forming at this point equal angles. We will call such a figure the Coxeter cross Σ_N (see Fig. 3). It is easy to construct a compactly supported function that is odd simultaneously with respect of all lines in Σ_N. Thus, a Coxeter cross is also a nonuniqueness set. The following result, conjectured in [60] and proven in the full generality in [7], shows that up to adding finitely many points, this is all that can happen to nonuniqueness sets:

Theorem 5 ([7]). *A set S in the plane $\mathbb{R}^2$ is a nonuniqueness set for compactly supported functions f, if and only if it belongs to the union $\Sigma_N \bigcup \Phi$ of a Coxeter cross Σ_N and a finite set of points Φ.*

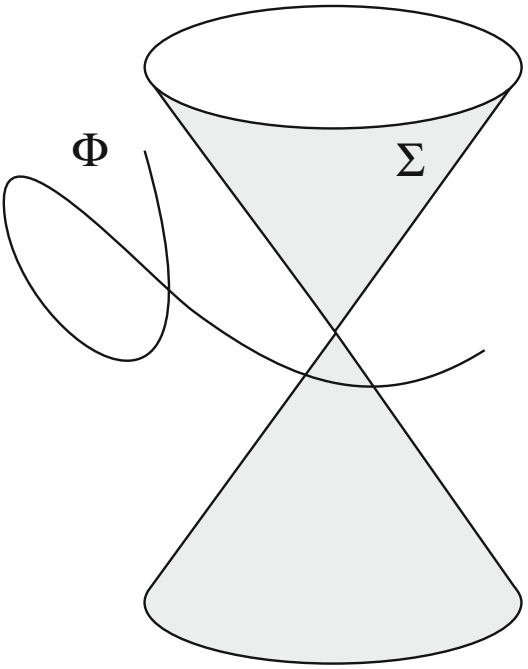


Fig. 4 The conjectured structure of a most general nonuniqueness set in $3D$

Again, compactness of support is crucial for the proof provided in [7]. There are no other proofs known at the moment of this result (see the corresponding open problem in Sect. 5). In particular, there is no proven analogue of Theorem 3 for non-closed sets S (unless S is an open part of a closed analytic surface).

The n-dimensional (in particular, $3D$) analogue of Theorem 5 has been conjectured [7], but never proven, although some partial advances in this direction have been made in [8, 36].

Conjecture 1. A set S in $\mathbb{R}^n$ is a nonuniqueness set for compactly supported functions f, if and only if it belongs to the union $\Sigma \bigcup \Phi$, where Σ is the cone of zeros of a homogeneous (with respect to some point in $\mathbb{R}^n$) harmonic polynomial, and Φ is an algebraic subset of $\mathbb{R}^n$ of dimension at most $n-2$ (see *Fig. 4*).

Uniqueness Results for a Finite Observation Time

So far, we have addressed only the question of uniqueness of reconstruction in the nonpractical case of the infinite observation time. There are, however, results that guarantee the uniqueness of reconstruction for a finite time of observation. The general idea is that it is sufficient to observe for the time that it takes the geometric rays (see section "Acoustically Inhomogeneous Media") from the interior Ω of S to reach S. In the case of a constant speed, which we will assume to be equal to 1,

the rays are straight and are traversed with the unit speed. This means that if D is the diameter of Ω (i.e., the maximal distance between two points in the closure of Ω), then after time $t = D$, all rays coming from Ω have left the domain. Thus, one hopes that waiting till time $t = D$ might be sufficient. In fact, due to the specific initial conditions in (3), namely, that the time derivative of the pressure is equal to zero at the initial moment, each singularity of f emanates two rays, and at least one of them will reach S in time not exceeding $D/2$. And indeed, the following result of [36] holds:

Theorem 6 ([36]). *If S is smooth and closed surface bounding domain Ω and D is the diameter of Ω, then the TAT data on S collected for the time $0 \le t \le 0.5D$ uniquely determines f.*

Notice that a shorter collection time does not guarantee uniqueness. Indeed, if S is a sphere and the observation time is less than $0.5D$, due to the finite speed of propagation, no information from a neighborhood of the center can reach S during observation. Thus, values of f in this neighborhood cannot be reconstructed.

Acoustically Inhomogeneous Media

We assume that the speed of sound is strictly positive, $c(x) > c > 0$, and such that $c(x) - 1$ has compact support, i.e., $c(x) = 1$ for large x.

Trapping and Non-trapping

We will frequently impose the so-called non-trapping condition on the speed of sound $c(x)$ in $\mathbb{R}^n$. To introduce it, let us consider the Hamiltonian system in $\mathbb{R}^{2n}_{x,\xi}$ with the Hamiltonian $H = \frac{c^2(x)}{2}|\xi|^2$:

$$\begin{cases} x_t' = \frac{\partial H}{\partial \xi} = c^2(x)\xi \\ \xi_t' = -\frac{\partial H}{\partial x} = -\frac{1}{2}\nabla\left(c^2(x)\right)|\xi|^2 \\ x|_{t=0} = x_0, \quad \xi|_{t=0} = \xi_0. \end{cases} \tag{10}$$

The solutions of this system are called bicharacteristics, and their projections into $\mathbb{R}^n_x$ are rays (or geometric rays).

Definition 3. We say that the speed of sound $c(x)$ satisfies the non-trapping condition, if all rays with $\xi_0 \neq 0$ tend to infinity when $t \to \infty$.

The rays that do not tend to infinity are called trapped.

A simple example, where quite a few rays are trapped, is the radial parabolic sound speed $c(x) = c|x|^2$.

It is well known (e.g., [46]) that singularities of solutions of the wave equation are carried by geometric rays. In order to make this statement more precise, we need

to recall the notion of a wave front set $WF(u)$ of a distribution $u(x)$ in $\mathbb{R}^n$. This set carries detailed information on singularities of $u(x)$.

Definition 4. Distribution $u(x)$ is said to be **microlocally smooth near a point** (x_0, ξ_0), where $x_0, \xi_0 \in \mathbb{R}^n$, and $\xi_0 \neq 0$, if there is a smooth "cutoff" function $\phi(x)$ such that $\phi(x_0) \neq 0$ and that the Fourier transform $\widehat{\phi u}(\xi)$ of the function $\phi(x)u(x)$ decays faster than any power $|\xi|^{-N}$ when $|\xi| \to \infty$, in directions that are close to the direction of ξ_0. (We remind the reader that if this Fourier transform decays that way in all directions, then $u(x)$ is smooth (infinitely differentiable) near the point x_0.)

The **wave front set** $WF(u) \subset \mathbb{R}^n_x \times \left(\mathbb{R}^n_\xi \backslash 0\right)$ **of** u consists of all pairs (x_0, ξ_0) such that u is not microlocally smooth near (x_0, ξ_0).

In other words, if $(x_0, \xi_0) \in WF(u)$, then u is not smooth near x_0, and the direction of ξ_0 indicates why it is not: the Fourier transform does not decay well in this direction. For instance, if $u(x)$ consists of two smooth pieces joined non-smoothly across a smooth interface Σ, then $WF(u)$ can only contain pairs (x, ξ) such that $x \in \Sigma$ and ξ is normal to Σ at x.

It is known that the wave front sets of solutions of the wave equation propagate with time along the bicharacteristics introduced above. This is a particular instance of a more general fact that applies to general PDEs and can be found in [46, 84]. As a result, if after time T all the rays leave the domain Ω of interest, the solution becomes smooth (infinitely differentiable) inside Ω.

The notion of so-called local energy decay, which we survey next, is important for the understanding of the non-trapping conditions in TAT.

Local Energy Decay Estimates

Assuming that the initial data $f(x)$ (1) is compactly supported and the speed $c(x)$ is non-trapping, one can provide the **local energy decay estimates** [32, 90, 91]. Namely, in any bounded domain Ω, the solution $p(x,t)$ of (1) satisfies, for a sufficiently large T_0 and for any (k, m), the estimate

$$\left| \frac{\partial^{k+|m|}}{\partial_t^k \partial_x^m} \right| \leq C_{k,m} \nu_k(t) \|f\|_{L^2}, \qquad \text{for } x \in \Omega, t > T_0. \tag{11}$$

Here $\nu_k(t) = t^{-n+1-k}$ for even n and $\nu_k(t) = e^{-\delta t}$ for odd n and some $\delta > 0$. Any value T_0 larger than the diameter of Ω works in this estimate.

Uniqueness Result for Non-trapping Speeds

If the speed is non-trapping, the local energy decay allows one to start solving the problem (3) from $t = \infty$, imposing zero conditions at $t = \infty$ and using the measured data g as the boundary conditions. This leads to recovery of the whole solution and in particular its initial value $f(x)$. As the result, one obtains the following simple uniqueness result of [3]:

Theorem 7 ([3]). *If the speed $c(x)$ is smooth and non-trapping and the acquisition surface S is closed, then the TAT data $g(y,t)$ determines the tomogram $f(x)$ uniquely.*

Notice that the statement of the theorem holds even if the support of f is not completely inside of the acquisition surface S.

Uniqueness Results for Finite Observation Times

As in the case of constant coefficients, if the speed of sound is non-trapping, appropriately long finite observation time suffices for the uniqueness. Let us denote by $T(\Omega)$ the *supremum* of the time it takes the ray to reach S, over all rays originating in Ω. In particular, if $c(x)$ is trapping, $T(\Omega)$ might be infinite.

Theorem 8 ([86]). *The data g measured till any time T larger than $T(\Omega)$ is sufficient for the unique recovery of f.*

Stability

By stability of reconstruction of the TAT tomogram f from the measured data g, we mean that small variations of g in an appropriate norm lead to small variations of the reconstructed tomogram f, also measured by an appropriate norm. In other words, small errors in the data lead to small errors in the reconstruction.

We will try to give the reader a feeling of the general state of affairs with stability, referring to the literature (e.g., [5, 48, 54, 71, 86]) for further exact details.

We will consider as functional spaces the standard Sobolev spaces H^s of smoothness s. We will also denote, as before, by $\mathcal{W}$ the operator transforming the unknown f into the data g.

Let us recall the notions of **Lipschitz and Hölder stability**. An even weaker **logarithmic stability** will not be addressed here. The reader can find discussion of the general stability notions and issues, as applied to inverse problems, in [49].

Definition 5. The operation of reconstructing f from g is said to be **Lipschitz stable** between the spaces H^{s_2} and H^{s_1}, if the following estimate holds for some constant C:

$$\|f\|_{H^{s_1}} \leq C\|g\|_{H^{s_2}}.$$

The reconstruction is said to be **Hölder stable** (a weaker concept), if there are constants $s_1, s_2, s_3, C, \mu > 0$, and $\delta > 0$ such that

$$\|f\|_{H^{s_1}} \leq C\|g\|^{\mu}_{H^{s_2}}$$

for all f such that $\|f\|_{H^{s_3}} \leq \delta$.

Stability can be also interpreted in the terms of the singular values σ_j of the forward operator $f \mapsto g$ in L^2, which have at most power decay when $j \to \infty$. The faster is the decay, the more unstable the reconstruction becomes. The problems with singular values decaying faster than any power of j are considered to be extremely unstable. Even worse are the problems with exponential decay of singular values (analytic continuation or solving Cauchy problem for an elliptic operator belong to this class). Again, the book [49] is a good source for finding detailed discussion of such issues.

Consider as an example inversion of the standard in X-ray CT and MRI Radon transform that integrates a function f over hyperplanes in $\mathbb{R}^n$. It smoothes function by "adding $(n-1)/2$ derivatives." Namely, it maps continuously H^s-functions in Ω into the Radon projections of class $H^{s+(n-1)/2}$. Moreover, the reconstruction procedure is Lipschitz stable between these spaces (see [63] for detailed discussion).

One should notice that since the forward mapping is smoothing (it "adds derivatives" to a function), the inversion should produce functions that are less smooth than the data, which is an unstable operation. The rule of thumb is that the stronger is the smoothing, the less stable is the inversion (this can be rigorously recast in the language of the decay of singular values). Thus, problems that require reconstructing non-smooth functions from infinitely differentiable (or even worse, analytic) data are extremely unstable (with super-algebraic or exponential decay of singular values correspondingly). This is just a consequence of the standard Sobolev embedding theorems (see, e.g., how this applies in TAT case in [65]).

In the case of a constant sound speed and the acquisition surface completely surrounding the object, as we have mentioned before, the TAT problem can be recast as inversion of the spherical mean transform $\mathcal{M}$ (see Sect. 2). Due to analogy between the spheres centered on S and hyperplanes, one can conjecture that the Lipschitz stability of **the inversion of the spherical mean operator $\mathcal{M}$ is similar to that of the inversion of the Radon transform**. This indeed is the case, **as long as f is supported inside S**, as has been shown in [71]. In the cases when closed-form inversion formulas are available (see section "Constant Speed of Sound"), this stability can also be extracted from them. If the support of f does reach outside, **reconstruction of the part of f that is outside is unstable** (i.e., is not even Hölder stable, due to the reasons explained in section "Incomplete Data").

In the case of **variable non-trapping speed of sound** $c(x)$, integral geometry does not apply anymore, and one needs to address the issue using, for instance, time reversal. In this case, stability follows by solving the wave equation in reverse time starting from $t = \infty$, as it is done in [3]. In fact, **Lipschitz stability in this case holds for any observation time exceeding** $T(\Omega)$ (see [86], where microlocal analysis is used to prove this result).

The bottom line is that **TAT reconstruction is sufficiently stable, as long as the speed of sound is non-trapping**.

However, trapping speed does cause instability [48]. Indeed, since some of the rays are trapped inside Ω, the information about some singularities never reaches S (no matter for how long one collects the data), and thus, as it is shown in [65], the reconstruction is not even Hölder stable between any Sobolev spaces, and the

singular values have super-algebraic decay. See also section "Incomplete Data" below for a related discussion.

Incomplete Data

In the standard X-ray CT, incompleteness of data arises, for instance, if not all projection angles are accessible or irradiation of certain regions is avoided, or as in the ROI (region of interest) imaging, only the ROI is irradiated.

It is not that clear what incomplete data means in TAT. Usually one says that one deals with **incomplete TAT data if the acquisition surface does not surround the object of imaging completely.** For instance, in breast imaging it is common that only a half-sphere arrangement of transducers is possible. We will see, however, that **incomplete data effects in TAT can also arise due to trapping, even if the acquisition surface completely surrounds the object**.

The questions addressed here are the following:

1. Is the collected incomplete data sufficient for **unique reconstruction**?
2. If yes, does the incompleteness of the data have any effect on the **stability and quality of the reconstruction**?

Uniqueness of Reconstruction

Uniqueness of reconstruction issues can be considered essentially resolved for incomplete data in TAT, at least in most situations of practical interest. We will briefly survey here some of the available results. In what follows, the acquisition surface S is not closed (otherwise the problem is considered to have complete data).

Uniqueness for Acoustically Homogeneous Media

In this case, Theorem 4 contains some useful sufficient conditions on S that guarantee uniqueness. Microlocal results of [7, 61, 85], as well as the PDE approach of [36] further applied in [8], provide also some other conditions. We assemble some of these in the following theorem:

Theorem 9. *Let S be a non-closed acquisition surface in TAT. Each of the following conditions on S is sufficient for the uniqueness of reconstruction of any compactly supported function f from the TAT data collected on S:*

1. *Surface S is not algebraic (i.e., there is no nonzero polynomial vanishing on S).*
2. *Surface S is a uniqueness set for harmonic polynomials (i.e., there is no nonzero harmonic polynomial vanishing on S).*
3. *Surface S contains an open piece of a closed analytic surface Γ.*
4. *Surface S contains an open piece of an analytic surface Γ separating the space $\mathbb{R}^n$ such that f is supported on one side of Γ.*
5. *For some point $y \in S$, the function f is supported on one side of the tangent plane T_y to S at y.*

For instance, if the acquisition surface S is just a tiny non-algebraic piece of a surface, data collected on S determines the tomogram f uniquely. However, one realizes that such data is unlikely to be useful for any practical reconstruction. Here the issue of stability of reconstruction kicks in, as it will be discussed in the stability subsection further down.

Uniqueness for Acoustically Inhomogeneous Media

In the case of a variable speed of sound, there still are uniqueness theorems for partial data [86, 87], e.g.:

Theorem 10 ([86]). *Let S be an open part of the boundary $\partial\Omega$ of a strictly convex domain Ω, and the smooth speed of sound equals 1 outside Ω. Then the TAT data collected on S for a time $T > T(\Omega)$ determines uniquely any function $f \in H_0^1(\Omega)$, whose support does not reach the boundary.*

A modification of this result that does not require strict convexity is also available in [87].

While useful uniqueness of reconstruction results exists for incomplete data problems, all such problems are expected to show instability. This issue is discussed in the subsections below. This will also lead to a better understanding of incomplete data phenomena in TAT.

"Visible" ("Audible") Singularities

According to the discussion in section "Acoustically Inhomogeneous Media," the singularities (the points of the wave front set $WF(f)$ of the function f in (3)) are transported with time along the bicharacteristics (10). Thus, in the x-space they are transported along the geometric rays. These rays may or may not reach the acquisition surface S, which triggers the introduction of the following notion:

Definition 6. A phase space point (x_0, ξ_0) is said to be "**visible**" (sometimes the word "**audible**" is used instead), if the corresponding ray (see (10)) reaches in finite time the observation surface S.

A region $U \subset \mathbb{R}^n$ is said to be in the **visibility zone**, if all points (x_0, ξ_0) with $x_0 \in U$ are visible.

An example of wave propagation through inhomogeneous medium is presented in Fig. 5. The open observation surface S in this example consists of the two horizontal and the left vertical sides of the square. Figure 5a shows some rays that bend, due to acoustic inhomogeneity, and leave through the opening of the observation surface S (the right side of the square). Figure 5b presents a flat phantom, whose wave front set creates these escaping rays, and thus is mostly invisible. Then Fig. 5c–f shows the propagation of the corresponding wave front.

Since the information about the horizontal boundaries of the phantom escapes, one does not expect to reconstruct it well. Figure 6 shows two phantoms and their reconstructions from the partial data: (a–b) correspond to the vertical flat

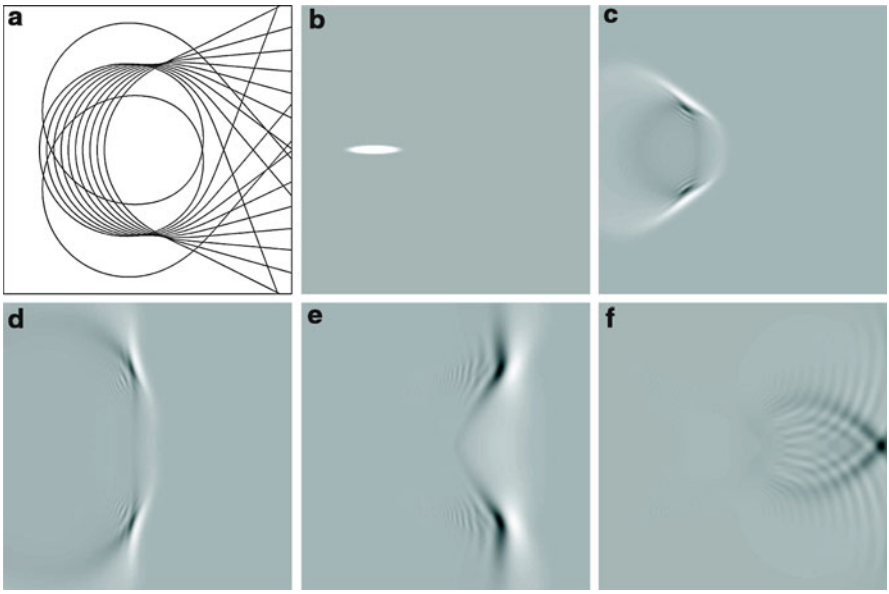


Fig. 5 (**a**) Some rays starting along the interval $x \in [-0.7, -0.2]$ in the vertical directions escape on the right; (**b**) a flat phantom with "invisible wave front"; (**c–f**) propagation of the flat front: most of the energy of the signal leaves the square domain through the hole on the right

phantom, whose only invisible singularities are at its ends. One sees essentially good reconstruction, with a little bit of blurring at the end points. On the other hand, reconstruction of the horizontal phantom with almost the whole wave front set invisible does not work. The next Fig. 7 shows a more complex square phantom, whose singularities corresponding to the horizontal boundaries are invisible, while the vertical boundaries are fine. One sees clearly that the invisible parts have been blurred away. On the other hand, Fig. 11a in Sect. 4 shows that one can reconstruct an image without blurring and with correct values, if the image is located in the visibility region. The reconstructed image in this figure is practically indistinguishable from the phantom shown in Fig. 10a.

Remark 2. If S is a closed surface and x_0 is a point outside of the convex hall of S, there is a vector $\xi_0 \neq 0$ such that (x_0, ξ_0) is "invisible." Thus, the visibility zone does not reach outside the closed acquisition surface S.

Stability of Reconstruction for Incomplete Data Problems

In all examples above, uniqueness of reconstruction held, but the images were still blurred. The question arises whether the blurring of "invisible" parts is avoidable (after all, the uniqueness theorems seem to claim that "everything is visible"). The answer to this is, in particular, the following result of [65], which is an analogue of similar statements in X-ray tomography:

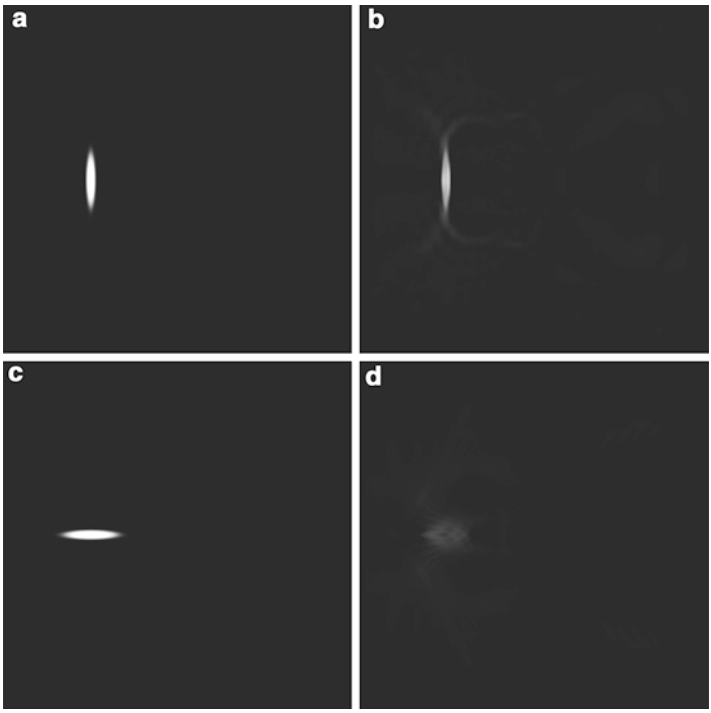


Fig. 6 Reconstruction with the same speed of sound as in Fig. 5: (**a**–**b**) a phantom with strong vertical fronts and its reconstruction; (**c**–**d**) a phantom with strong horizontal fronts and its reconstruction

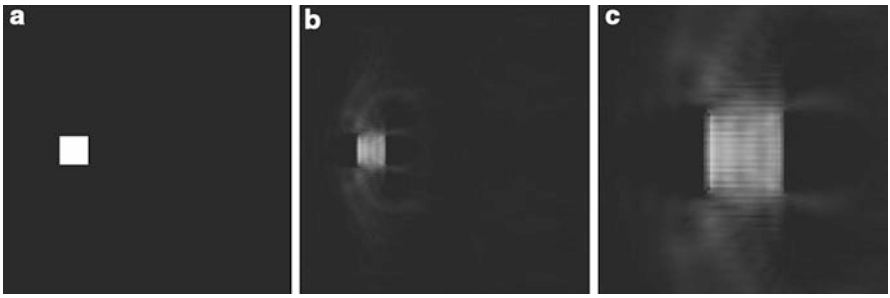


Fig. 7 Reconstruction with the same speed of sound as in Fig. 5: (**a**) a phantom; (**b**) its reconstruction; (**c**) a magnified fragment of (**b**)

Theorem 11 ([65]). *If there are invisible points* (x_0, ξ_0) *in* $\Omega \times \left(\mathbb{R}^n_\xi \backslash 0\right)$, *then inversion of the forward operator* $\mathcal{W}$ *is not Hölder stable in any Sobolev spaces. The singular values* σ_j *of* $\mathcal{W}$ *in* L^2 *decay super-algebraically.*

Thus, the presence of invisible singularities makes the reconstruction severely ill-posed. In particular, according to Remark 2, this theorem implies the following statement:

Corollary 1. *Reconstruction of the parts of $f(x)$ supported outside the closed observation surface S is unstable.*

On the other hand,

Theorem 12 ([86]). *All visible singularities of f can be reconstructed with Lipschitz stability (in appropriate spaces).*

Such a reconstruction of visible singularities can be obtained in many ways, for instance, just by replacing the missing data by zeros (with some smoothing along the junctions with the known data, in order to avoid artifact singularities). However, there is no hope for stable recovery of the correct values of $f(x)$, if there are invisible singularities.

Discussion of the Visibility Condition

Visibility for Acoustically Homogeneous Media

In the constant speed case, the rays are straight, and thus the visibility condition has a simple test:

Proposition 1 (e.g., [48, 99, 100]). *If the speed is constant, a point x_0 is in the visible region, if and only if any line passing through x_0 intersects at least once the acquisition surface S (and thus a detector location).*

Figure 8 illustrates this statement. It shows a square phantom and its reconstruction from complete data and from the data collected on the half circle S surrounding the left half of object. The parts of the interfaces where the normal to the interface that does not cross S are blurred.

Visibility for Acoustically Inhomogeneous Media

When the speed of sound is variable, an analogue of Proposition 1 holds, with lines replaced by rays.

Proposition 2 (e.g., [48, 65, 86]). *A point x_0 is in the visible region, if and only if for any $\xi_0 \neq 0$ at least one of the two geometric rays starting at (x_0, ξ_0) and at $(x_0, -\xi_0)$ (see (10)) intersects the acquisition surface S (and thus a detector location).*

The reader can now see an important difference between the acoustically homogeneous and inhomogeneous media. Indeed, even if S surrounds the support

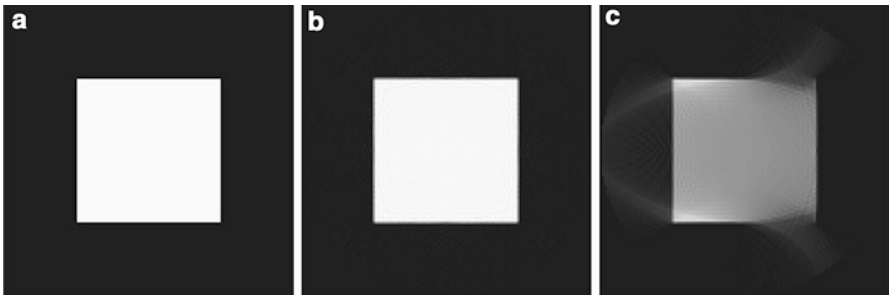


Fig. 8 Reconstruction from incomplete data using closed-form inversion formula in $2D$; detectors are located on the left half circle of radius 1.05: (**a**) phantom; (**b**) reconstruction from complete data; and (**c**) reconstruction from the incomplete data

of f completely, trapped rays will never find their way to S, which will lead, as we know by now, to instabilities and blurring of some interfaces.

Thus, the presence of rays trapped inside the acquisition surface creates effects of incomplete data type. This is exemplified in Fig. 9 with a square phantom and its reconstruction shown in the presence of a trapping (parabolic) speed. Notice that the square centered at the center of symmetry of the speed is reconstructed very well (see Fig. 9d), since none of the rays carrying its singularities is trapped.

Range Conditions

In this section we address the problem of describing the ranges of the forward operators $\mathcal{W}$ (see (4)) and $\mathcal{M}$ (see (8)), the latter in the case of an acoustically homogeneous medium (i.e., for $c = \text{const}$). The ranges of these operators, similarly to the range of the Radon and X-ray transforms (see [63]), are of infinite co-dimensions. This means that ideal data g from a suitable function space satisfy infinitely many mandatory identities. Knowing the range is useful for many theoretical and practical purposes in various types of tomography (reconstruction algorithms, error corrections, incomplete data completion, etc.), and thus this topic has attracted a lot of attention (e.g., [41, 45, 63, 68, 70] and references therein).

As we will see in the next section, range descriptions in TAT are also intimately related to the recovery of the unknown speed of sound.

We recall [41, 45, 63] that for the standard Radon transform

$$f(x) \to g(s,\omega) = \int\limits_{x\cdot\omega=s} f(x)dx, |\omega| = 1,$$

where f is assumed to be smooth and supported in the unit ball $B = \{x \mid |x| \le 1\}$, the range conditions on $g(s,\omega)$ are:

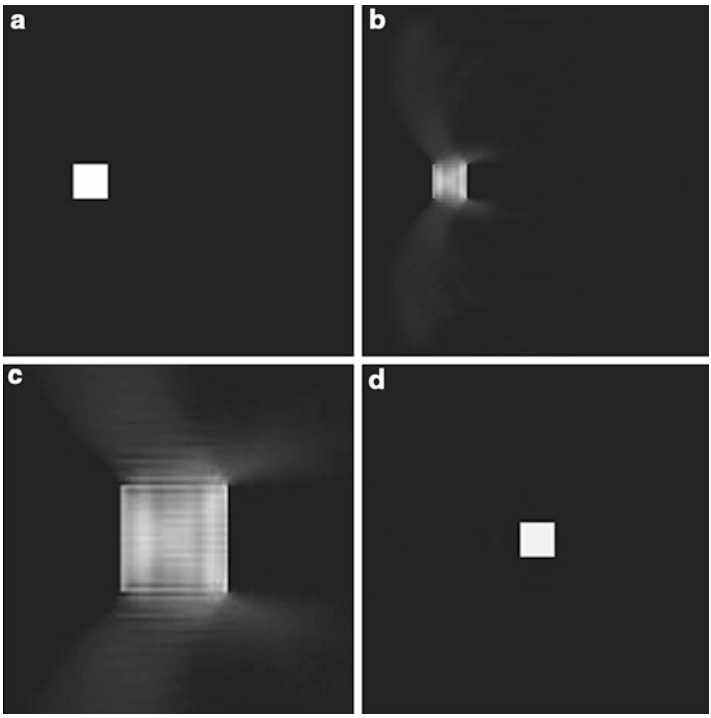


Fig. 9 Reconstruction of a square phantom from full data in the presence of a trapping parabolic speed of sound (the speed is radial with respect to the center of the picture): (**a**) an off-center phantom; (**b**) its reconstruction; (**c**) a magnified fragment of (**b**); (**d**) reconstruction of a centered square phantom

1. *Smoothness and support*: $g \in C_0^\infty([-1,1] \times \mathcal{S})$, where $\mathcal{S}$ is the unit sphere of vectors ω
2. *Evenness*: $g(-s,-\omega) = g(s,\omega)$
3. *Moment conditions*: for any integer $k \geq 0$, the kth moment

$$G_k(\omega) = \int_{-\infty}^{\infty} s^k g(\omega, s) ds$$

extends from the unit sphere $\mathcal{S}$ to a homogeneous polynomial of degree k in ω.

The seemingly "trivial" evenness condition is sometimes the hardest to generalize to other transforms of Radon type, while it is often easier to find analogues of the moment conditions. This is exactly what happens in TAT.

For the operators $\mathcal{W}, \mathcal{M}$ in TAT, some sets of range conditions of the moment type had been discovered over the years [7,60,78], but complete range descriptions started to emerge only since 2006 [2,4–6,9,37,54].

Range descriptions for the more general operator $\mathcal{W}$ are harder to obtain than for $\mathcal{M}$, and complete range descriptions are not known for even dimensions or for the case of the variable speed of sound.

Let us address the case of the spherical mean operator $\mathcal{M}$ first.

The Range of the Spherical Mean Operator $\mathcal{M}$

The support and smoothness conditions are not hard to come up with, at least when S is a sphere. By choosing an appropriate length scale, we can assume that the sphere is of radius 1 and centered at the origin and that the interior domain Ω is the unit ball $B = \{x \mid |x| = 1\}$. If f is smooth and supported inside B (i.e., $f \in C_0^\infty(B)$), then it is clear that the measured data satisfies the following:

Smoothness and support conditions:

$$g \in C_0^\infty(S \times [0,2]). \tag{12}$$

An analogue of the moment conditions for $g(y,r) := \mathcal{M}f$ was implicitly present in [7,60] and explicitly formulated as such in [78]:

Moment conditions: *for any integer $k \geq 0$, the moment*

$$M_k(y) = \int_0^\infty r^{2k+d-1} g(y,r) dr \tag{13}$$

extends from S to a (in general, nonhomogeneous) polynomial $Q_k(x)$ of degree at most $2k$.

These two types of conditions happen to be incomplete, i.e., infinitely many others exist. The Radon transform experience suggests to look for an analogue of evenness conditions. And indeed, a set of conditions called orthogonality conditions was found in [5,9,37].

Orthogonality conditions: Let $-\lambda_k^2$ be the eigenvalue of the Laplace operator Δ in B with zero Dirichlet conditions and ψ_k be the corresponding eigenfunctions. Then the following orthogonality condition is satisfied:

$$\int_{S\times[0,2]} g(x,t)\partial_\nu \psi_\lambda(x) j_{n/2-1}(\lambda t) t^{n-1} dx dt = 0. \tag{14}$$

Here $j_p(z) = c_p z^{-p} J_p(z)$ is the so-called spherical Bessel function.

The range descriptions obtained in [5,9,37] demonstrated that these three types of conditions completely describe the range of the operator $\mathcal{M}$ on functions $f \in C_0^\infty(B)$. At the same time, the results of [5,37] showed that the moment conditions can be dropped in odd dimensions. It was then discovered in [2] that the moment

conditions can be dropped altogether in any dimension, since they follow from the other two types of conditions:

Theorem 13 ([2]). *Let S be the unit sphere. A function $g(y,t)$ on the cylinder $S \times \mathbb{R}^+$ can be represented as $\mathcal{M}f$ for some $f \in C_0^\infty(B)$ if an only if it satisfied the above smoothness and support and orthogonality conditions* (12), (14).

The statement also holds in the finite smoothness case, if one replaces the requirements by $f \in H_0^s(B)$ and $g \in H_0^{s+(n-1)/2}(S \times [0,2])$.

The range of the forward operator $\mathcal{M}$ has not been described when S is not a sphere, but, say, a convex smooth closed surface. The moment and orthogonality conditions hold for any S, and appropriate smoothness and support conditions can also be formulated, at least in the convex case. However, it has not been proven that they provide the complete range description.

It is quite possible that for nonspherical S the moment conditions might have to be included into the range description.

A different range description of the Fredholm alternative type was developed in [71] (see also [39] for description of this result).

The Range of the Forward Operator $\mathcal{W}$

We recall that the operator $\mathcal{W}$ (see (4)) transforms the initial value f in (3) into the observed S values g of the solution. There exist Kirchhoff–Poisson formulas representing the solution p and thus $g = \mathcal{W}f$ in terms of the spherical means of f (i.e., in terms of $\mathcal{M}f$). However, translating the result of Theorem 13 into the language of $\mathcal{W}$ is not straightforward, since in even dimensions these formulas are nonlocal ([27] p. 682):

$$\mathcal{W}f(y,t) = \frac{\sqrt{\pi}}{2\Gamma(n/2)} \left(\frac{1}{t}\frac{\partial}{\partial t}\right)^{(n-3)/2} t^{n-2} \left(\mathcal{M}f\right)(y,t), \text{ for odd } n. \tag{15}$$

and

$$\mathcal{W}f(y,t) = \frac{1}{\Gamma(n/2)} \left(\frac{1}{t}\frac{\partial}{\partial t}\right)^{(n-2)/2} \int_0^t \frac{r^{n-1}\left(\mathcal{M}f\right)(y,r)}{\sqrt{t^2-r^2}} dr, \text{ for even } n. \tag{16}$$

The nonlocality of the transformation for even dimensions reflects the absence of Huygens' principle (i.e., absence of sharp rear fronts of waves) in these dimensions; it also causes difficulties in establishing the complete range descriptions. In particular, due to the integration in (16), $\mathcal{M}f(y,t)$ does not vanish for large times t anymore. One can try to use other known operators intertwining the two problems (see [5] and references therein), and some of which do preserve vanishing for large values of t, but this so far has led only to very clumsy range descriptions.

However, for odd dimensions, the range description of $\mathcal{W}$ can be obtained. In order to do so, given the TAT data $g(y,t)$, let us introduce an auxiliary time-reversed problem in the cylinder $B \times [0,2]$:

$$\begin{cases} q_{tt} - \Delta q = 0 \text{ for } (x,t) \in B \times [0,2]), \\ q(x,2) = q_t(x,2) = 0 \text{ for } x \in B, \\ q(y,t) = g(y,t) \text{ for } (y,t) \in S \times [0,2]). \end{cases} \tag{17}$$

We can now formulate the range description from [37, 39]:

Theorem 14 ([37, 39]). *For odd dimensions n and S being the unit sphere, a function $g \in C_0^\infty(S \times [0,2])$ can be represented as $\mathcal{W}f$ for some $f \in C_0^\infty(B)$ if and only if the following condition is satisfied:*

The solution q of (17) satisfies $q_t(x,0) = 0$ for all $x \in B$.

Orthogonality type and Fredholm alternative type range conditions, equivalent to the one in the theorem above, are also provided in [37, 39].

Reconstruction of the Speed of Sound

Unsurprisingly, all inversion procedures outlined in Sect. 4 rely upon the knowledge of the speed of sound $c(x)$. Although often, e.g., in breast imaging, the medium is assumed to be acoustically homogeneous, this is not a good assumption in many other cases. It has been observed (e.g., [48, 50]) that replacing even slightly varying speed of sound with its average value might significantly distort the image; not only the numerical values, but also the shapes of interfaces between the tissues will be reconstructed incorrectly. Thus, the question of estimating $c(x)$ correctly becomes important. One possible approach [50] is to use an additional transmission ultrasound scan to reconstruct the speed beforehand. The question arises of whether one could determine the speed of sound $c(x)$ and the tomogram $f(x)$ (assuming that f is not zero) simultaneously from the TAT data. In fact, one needs only to determine $c(x)$ (without knowing f), since then inversion procedures of Sect. 4 would apply to recover f.

At the first glance, this seems to be an overly ambitious project. Indeed, if we denote the forward operator $\mathcal{W}$ by $\mathcal{W}_c$, to indicate its dependence on the speed of sound $c(x)$, then the problem becomes, given the data g, to find both c and f from the equality

$$\mathcal{W}_c f = g. \tag{18}$$

A similar situation arises in the SPECT emission tomography (see [63] and references therein), where the role of the speed of sound is played by the unknown attenuation. It is known, however, that in SPECT the attenuation can be recovered for a "generic" f.

What is the reason for such a strange situation? It looks like for any c one could solve Eq. (18) for an f, and thus no information about c is contained in the data

g. This argument is incorrect for the following reason: the range of the forward operator, as we know already from the previous section, has infinite co-dimension. Thus, this range has a lot of space to "rotate" when c changes. Imagine for an instance that the rotation is so powerful that for different values of c the ranges have only zero (the origin) in common. Then, knowing g in the range, one would know which c it came from. Thus, the problem of recovering the speed of sound from the TAT data is closely related to the range descriptions.

Numerical inversions using algebraic iterative techniques (e.g., [102, 104]) show that recovering both c and f might be indeed possible.

Unfortunately, very little is known at the moment concerning this problem. Direct usage of range conditions attempted in [48] has led only to extremely weak and not practically useful results so far. A revealing relation to the transmission eigenvalue problem well known in inverse problems (see [26] for the survey) was recently discovered by D. Finch. Unfortunately, the transmission eigenvalue problem remains still unresolved. However, one can derive from this relation the following result regarding uniqueness of the reconstruction of the speed of sound, due to M. Agranovsky (a somewhat restricted version is due to D. Finch et al., both unpublished):

Theorem 15. *If two speeds satisfy the inequality* $c_1(x) \geq c_2(x)$ *for all* $x \in \Omega$ *and produce for some functions* f_1, f_2 *the same nonzero TAT data* g *(i.e.,* $\mathcal{W}c_1 f_1 = g, \mathcal{W}c_2 f_2 = g$*), then* $c_1(x) = c_2(x)$.

It is known [49, Corollary 8.2.3] that if a function $f(x)$ is such that $\Delta f(x) \neq 0$ and for two acoustic speeds $c_1(x)$ and $c_2(x)$, it produces the same TAT data g, then $c_1 = c_2$.

It is clear that the problem of finding the speed of sound from the TAT data is still mostly unresolved.

4 Reconstruction Formulas, Numerical Methods, and Case Examples

Numerous formulas, algorithms, and procedures for reconstruction of images from TAT measurements have been developed by now. Most of these techniques require the data being collected on a closed surface (closed curve in $2D$) surrounding the object to be imaged. Such methods are discussed in section "Full Data (Closed Acquisition Surfaces)." We review methods that work under the assumption of constant speed of sound in section "Constant Speed of Sound." The techniques applicable in the case of the known variable speed of sound are considered in section "Constant Speed of Sound." Closed surface measurements cannot always be implemented, since in some practical situations the object cannot be completely surrounded by the detectors. In this case, one has to resort to various approximate reconstruction techniques as discussed in section "Partial (Incomplete) Data."

Full Data (Closed Acquisition Surfaces)

Constant Speed of Sound

When the speed of sound within the tissues is a known constant, the TAT problem can be reformulated (see Sect. 2) in terms of the values of the spherical means of the initial condition $f(x)$. These means can be easily recovered from the measurements of the acoustic pressure using formulas (15) and (16) (see the discussion in [7]). In this case, image reconstruction becomes equivalent to inverting the spherical mean transform $\mathcal{M}$. Thus, in what follows, we consider the problem of reconstructing a function $f(x)$ supported within the region bounded by a closed surface S from known values of its spherical integrals $g(y, r)$ with centers on S

$$g(y,r) = \int\limits_{\mathbb{S}^{n-1}} f(y + r\omega) r^{n-1} d\omega, \quad y \in S, \tag{19}$$

where $d\omega$ is the standard measure on the unit sphere.

Series Solutions for Spherical Geometry

The first inversion procedures for the case of closed acquisition surfaces were described in [66, 67], where solutions were found for the cases of circular (in $2D$) and spherical (in $3D$) surfaces, respectively. These solutions were obtained by the harmonic decomposition of the measured data and of the sought function $f(x)$, followed by equating coefficients of the corresponding Fourier series. In particular, the $2D$ algorithm of [66] pertains to the case when the detectors are located on a circle of radius R. This method is based on the Fourier decomposition of f and g in angular variables

$$\begin{aligned} f(x) &= \sum_{-\infty}^{\infty} f_k(\rho) e^{ik\varphi}, \quad x = (\rho\cos(\varphi), \rho\sin(\varphi)) \\ g(y(\theta), r) &= \sum_{-\infty}^{\infty} g_k(r) e^{ik\theta}, \quad y = (R\cos(\theta), R\sin(\theta)), \end{aligned} \tag{20}$$

where

$$(\mathcal{H}_m u)(s) = 2\pi \int_0^{\infty} u(t) J_m(st) t\, dt$$

is the Hankel transform and $J_m(t)$ is the Bessel function. As shown in [66], the Fourier coefficients $f_k(\rho)$ can be recovered from the known coefficients $g_k(r)$ by the following formula:

$$f_k(\rho) = \mathcal{H}_m \left(\frac{1}{J_k(\lambda|R|)} \mathcal{H}_0 \left[\frac{g_k(r)}{2\pi r} \right] \right).$$

This method requires division of the Hankel transform of the measured data by the Bessel functions J_k, which have infinitely many zeros. Theoretically, there is no problem: the range conditions (section "Range Conditions") on the exact data g imply that the Hankel transform $\mathcal{H}_0\left[(2\pi r)^{-1} g_k(r)\right]$ has zeros that cancel those in the denominator. However, since the measured data always contain errors, the exact cancelation does not happen, and one needs a sophisticated regularization scheme to guarantee that the error remains bounded.

This difficulty can be avoided (see, e.g., [54]) by replacing the Bessel function J_0 in the inner Hankel transform by the Hankel function $H_0^{(1)}$. This yields the following formula for $f_k(\rho)$:

$$f_k(\rho) = \mathcal{H}_k\left(\frac{1}{H_k^{(1)}(\lambda|R|)}\int_0^\infty g_k(r) H_0^{(1)}(\lambda r) dr\right).$$

Unlike J_m, Hankel functions $H_m^{(1)}(t)$ do not have zeros for any real values of t, which removes the problems with division by zeros [66]. (A different way of avoiding divisions by zero was found in [44].)

This derivation can be repeated in $3D$, with the exponentials $e^{ik\theta}$ replaced by the spherical harmonics and with cylindrical Bessel functions replaced by their spherical counterparts. By doing this, one arrives at the Fourier series method of [67] (see also [97]). The use of the Hankel function $H_0^{(1)}$ above is similar to the way the spherical Hankel function $h_0^{(1)}$ is utilized in [67] to avoid the divisions by zero.

Eigenfunction Expansions for a General Geometry

The series methods described in the previous section rely on the separation of variables that occurs only in spherical geometry. A different approach was proposed in [58]. It works for arbitrary closed surfaces, but is practical only for those with explicitly known eigenvalues and eigenfunctions of the Dirichlet Laplacian in the interior. These include, in particular, the surfaces of such bodies as spheres, half spheres, cylinders, cubes, and parallelepipeds, as well as the surfaces of crystallographic domains.

Let λ_m^2 and $u_m(x)$ be the eigenvalues and an orthonormal basis of eigenfunctions of the Dirichlet Laplacian $-\Delta$ in the interior Ω of a closed surface S:

$$\begin{aligned} \Delta u_m(x) + \lambda_m^2 u_m(x) &= 0, \quad x \in \Omega, \quad \Omega \subseteq \mathbb{R}^n, \\ u_m(x) &= 0, \quad x \in S, \\ ||u_m||_2^2 &\equiv \int_\Omega |u_m(x)|^2 dx = 1. \end{aligned} \tag{21}$$

As before, one would like to reconstruct a compactly supported function $f(x)$ from the known values of its spherical integrals $g(y, r)$ (see (19)) with centers on S. Since $u_m(x)$ is the solution of the Dirichlet problem for the Helmholtz equation with zero

boundary conditions and the wave number λ_m, this function admits the Helmholtz representation

$$u_m(x) = \int_S \Phi_{\lambda_m}(|x-y|)\frac{\partial}{\partial n}u_m(y)ds(y) \quad x \in \Omega, \tag{22}$$

where $\Phi_{\lambda_m}(|x-y|)$ is a free-space Green's function of the Helmholtz equation (21) and n is the exterior normal to S.

The function $f(x)$ can be expanded into the series

$$f(x) = \sum_{m=0}^{\infty} \alpha_m u_m(x), \text{ where} \tag{23}$$

$$\alpha_m = \int_\Omega u_m(x) f(x) dx.$$

A reconstruction formula for α_m (and thus for $f(x)$) will result, if one substitutes representation (22) into (23) and interchanges the orders of integration

$$\alpha_m = \int_\Omega u_m(x) f(x) dx = \int_S I(y, \lambda_m)\frac{\partial}{\partial n}u_m(y) dA(x), \tag{24}$$

where

$$I(y,\lambda) = \int_\Omega \Phi_\lambda(|x-y|) f(x) dx = \int_0^{\operatorname{diam}\Omega} g(y,r)\Phi_\lambda(r) dr. \tag{25}$$

Now $f(x)$ can be obtained by summing the series (23). This method becomes computationally efficient when the eigenvalues and eigenfunctions are known explicitly, especially if a fast summation formula for the series (23) is available. This is the case when the acquisition surface S is the surface of a cube, and thus the eigenfunctions are products of sine functions. The resulting $3D$ reconstruction algorithm is extremely fast and precise (see [58]).

The above method has an interesting property. If the support of the source $f(x)$ extends outside Ω, the algorithm still yields theoretically an exact reconstruction of $f(x)$ inside Ω. Indeed, the value of the expression (22) for all x lying outside Ω is zero. Thus, when one computes (24) for $x \in \mathbb{R}^n \backslash \Omega$, values of $f(x)$ are multiplied by zero and do not affect further computation in any way. This feature is shared by the time-reversal method (see the corresponding paragraph in section "Constant Speed of Sound"). The closed-form FBP-type reconstruction techniques considered in the next subsection do not have this property. In other words, in the presence of a source outside the measurement surface, reconstruction within Ω can be incorrect.

The reason for this difference is that all currently known closed-form FBP-type formulas rely (implicitly or explicitly) on the assumption that the wave propagates outside S in the whole free space and has no sources outside. On the other hand,

the eigenfunction expansion method and the time reversal rely only upon the time decay of the wave inside S, which is not influenced by f having a part outside S.

Closed-Form Inversion Formulas

Closed-form inversion formulas play a special role in tomography. They bring about better theoretical understanding of the problem and frequently serve as starting points for the development of efficient reconstruction algorithms. A well-known example of the use of explicit inversion formulas is the so-called filtered backprojection (FBP) algorithm in X-ray tomography, which is derived from one of the inversion formulas for the classical Radon transform (see, e.g., [63]).

The very existence of closed-form inversion formulas for TAT had been in doubt, till the first such formulas were obtained in odd dimensions by Finch et al. in [36], under the assumption that the acquisition surface S is a sphere. Suppose that the function $f(x)$ is supported within a ball of radius R and that the detectors are located on the surface $S = \partial B$ of this ball. Then some of the formulas obtained in [36] read as follows:

$$f(x) = -\frac{1}{8\pi^2 R}\Delta_x \int\limits_{\partial B} \frac{g(y, |y-x|)}{|y-x|} dA(y), \tag{26}$$

$$f(x) = -\frac{1}{8\pi^2 R} \int\limits_{\partial B} \left(\frac{1}{r}\frac{\partial^2}{\partial r^2} g(y,r)\right)\Bigg|_{r=|y-x|} dA(y), \tag{27}$$

$$f(x) = -\frac{1}{8\pi^2 R} \int\limits_{\partial B} \left(\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial}{\partial r}\frac{g(y,r)}{r}\right)\right)\Bigg|_{r=|y-x|} dA(y), \tag{28}$$

where $dA(y)$ is the surface measure on ∂B and g represents the values of the spherical integrals (19).

These formulas have an FBP (filtered backprojection) nature. Indeed, differentiation with respect to r in (27) and (28) and the Laplace operator in (26) represent the filtration, while the (weighted) integrals correspond to the backprojection, i.e., integration over the set of spheres passing through the point of interest x and centered on S.

The so-called universal backprojection formula in $3D$ was found in [98] (it is also valid for the cylindrical and plane acquisition surfaces, see section "Partial (Incomplete) Data"). In our notation, this formula takes the form

$$f(x) = \frac{1}{8\pi^2}\mathrm{div} \int\limits_{\partial B} n(y) \left(\frac{1}{r}\frac{\partial}{\partial r}\frac{g(y,r)}{r}\right)\Bigg|_{r=|y-x|} dA(y), \tag{29}$$

or, equivalently,

$$f(x) = -\frac{1}{8\pi^2}\int\limits_{\partial B} \frac{\partial}{\partial n}\left(\frac{1}{r}\frac{\partial}{\partial r}\frac{g(y,r)}{r}\right)\Bigg|_{r=|y-x|} dA(y), \tag{30}$$

where $n(y)$ is the exterior normal vector to ∂B. One can show [4, 64, 98] that formulas (26) through (29) are not equivalent on non-perfect data: the result will differ if these formulas are applied to a function that does not belong to the range of the spherical mean transform $\mathcal{M}$. A family of inversion formulas valid in $\mathbb{R}^n$ for arbitrary $n \geq 2$ was found in [57]

$$f(x) = \frac{1}{4(2\pi)^{n-1}}\text{div}\int\limits_{\partial B} n(y)h(y,|x-y|)dA(y), \tag{31}$$

where

$$h(y,t) = \int\limits_{\mathbb{R}^+} Y(\lambda t)\left[\int\limits_0^{2R} J(\lambda r)g(y,r)dr - J(\lambda t)\int\limits_0^{2R} Y(\lambda r)g(y,r)dr\right]\lambda^{2n-3}d\lambda, \tag{32}$$

$$J(t) = \frac{J_{n/2-1}(t)}{t^{n/2-1}}, \qquad Y(t) = \frac{Y_{n/2-1}(t)}{t^{n/2-1}}, \tag{33}$$

and $J_{n/2-1}(t)$ and $Y_{n/2-1}(t)$ are, respectively, the Bessel and Neumann functions of order $n/2-1$. In $3D$, $J(t)$ and $Y(t)$ are simply $t^{-1}\sin t$ and $t^{-1}\cos t$, and formulas (31) and (32) reduce to (30).

In $2D$, Eq. (32) also can be simplified [4], which results in the formula

$$f(x) = \frac{1}{2\pi^2}\text{div}\int\limits_{\partial B} n(y)\left[\int\limits_0^{2R} g(y,r)\frac{1}{r^2-|x-y|^2}dr\right]dl(y), \tag{34}$$

where ∂B now stands for the circle of radius R and $dl(y)$ is the standard arc length.

A different set of closed-form inversion formulas applicable in even dimensions was found in [35]. Formula (34) can be compared to the following inversion formulas from [35]

$$f(x) = \frac{1}{2\pi R}\Delta\int\limits_{\partial B}\int\limits_0^{2R} g(y,r)\log(r^2-|x-y|^2)\,dr\,dl(y), \tag{35}$$

or

$$f(x) = \frac{1}{2\pi R}\int\limits_{\partial B}\int\limits_0^{2R}\frac{\partial}{\partial r}\left(r\frac{\partial}{\partial r}\frac{g(y,r)}{r}\right)\log(r^2-|x-y|^2)\,dr\,dl(y). \tag{36}$$

Finally, a unified family of inversion formulas was derived in [64]. In our notation, it has the following form

$$f(x) = -\frac{4}{\pi R}\int\limits_{\partial B}\left(\frac{\partial}{\partial t}K_n(y,t)\right)\Bigg|_{t=|x-y|}\frac{<y-x,y-\xi>}{|x-y|}dA(y), \quad (37)$$

$$K_n(y,t) = -\frac{1}{16(2\pi)^{n-2}}\int\limits_{\mathbb{R}^+}\lambda^{2n-3}Y(\lambda t)\left(\int\limits_{\mathbb{R}^+}J(\lambda r)g(y,r)dr\right)d\lambda$$

where ∂B is the surface of a ball in $\mathbb{R}^n$ of radius R, functions J and Y are in (33), and ξ is an arbitrary fixed vector. In particular, in $3D$

$$J(t) = \sqrt{\frac{2}{\pi}}\frac{\sin t}{t}, J(t) = \sqrt{\frac{2}{\pi}}\frac{\cos t}{t}$$

and, after simple calculation, the above inversion formula reduces to

$$f(x) = -\frac{1}{8\pi^2 R}\int\limits_{\partial B}\left(\frac{\partial}{\partial r}\frac{1}{r}\frac{\partial}{\partial r}\frac{g(y,r)}{r}\right)\Bigg|_{r=|x-y|}\frac{<y-x,y-\xi>}{|x-y|}dA(y). \quad (38)$$

Different choices of vector ξ in the above formula result in different inversion formulas. For example, if ξ is set to zero, the ratio $\frac{<y-x,y-\xi>}{|x-y|}$ equals $R\cos\alpha$, where α is the angle between the exterior normal $n(y)$ and the vector $y-x$; when combined with the derivative in t, this factor produces the normal derivative, and the inversion formula (38) reduces to (30). On the other hand, the choice of $\xi = x$ in (38) leads to a formula

$$f(x) = -\frac{1}{8\pi^2 R}\int\limits_{\partial B}\left(r\frac{\partial}{\partial r}\frac{1}{r}\frac{\partial}{\partial r}\frac{g(y,r)}{r}\right)\Bigg|_{r=|x-y|}dA(y),$$

which is reminiscent of formulas (26)–(28).

Green's Formula Approach and Some Symmetry Considerations

Let us suppose for a moment that the acoustic detectors could measure not only the pressure $p(y,t)$ at each point of the acquisition surface S, but also the normal derivative $\partial p/\partial n$ on S. Then the problem of reconstructing the initial pressure $f(x)$ becomes rather simple. Indeed, one can use the knowledge of the free-space Green's function for the wave equation and invoke the Green's theorem to represent the solution $p(x,t)$ of (3) in the form of integrals over S involving $p(x,t)$ and its

normal derivative and the Green's function and its normal derivative. (This can be done in the Fourier or time domains.) This would require infinite observation time, but in $3D$ the time $T(\Omega)$ will suffice, after which the wave escapes the region of interest (a cutoff also would work approximately in $2D$ similar to the time-reversal method). This Green's function approach happens to be, explicitly or implicitly, the starting point of all closed-form inversions described above. The trick is to rewrite the formula in such a way that the unknown in reality normal derivative $\partial p/\partial n$ disappears from the formula.

This was achieved in [57] by reducing the question to some integrals involving special functions and making the key observation that the integral

$$I_\lambda(x,y) = \int\limits_{\partial B} J(\lambda|x-z|)\frac{\partial}{\partial n}Y(\lambda|y-z|)dA(z), \quad x,y \in B \subset \mathbb{R}^n$$

is a symmetric function of its arguments:

$$I_\lambda(x,y) = I_\lambda(y,x) \text{ for } x,y \in B \subset \mathbb{R}.^n \tag{39}$$

Similarly, the derivation of (37) in [64] employs the symmetry of the integral

$$K_\lambda(x,y) = \int\limits_{\partial B} J(\lambda|x-z|)Y(\lambda|y-z|)dA(z), \quad x,y \in B \subset \mathbb{R}^n.$$

In fact, the symmetry holds for any integral

$$W_\lambda(x,y) = \int\limits_{\partial B} U(\lambda|x-z|)V(\lambda|y-z|)dA(z), \quad x,y \in B \subset \mathbb{R}^n,$$

where $U(\lambda|x|)$ and $V(\lambda|x|)$ are any two radial solutions of the Helmholtz equation

$$\Delta u(x) + \lambda^2 u(x) = 0. \tag{40}$$

It is straightforward to verify this symmetry when S is a sphere and B is the corresponding ball, and the points x, y lie on the boundary S only, rather than anywhere in B. This follows immediately from the rotational symmetry of S. The same is true for the normal derivatives on S of $W_\lambda(x,y)$ in x and y.

This boundary symmetry happens to imply the needed full symmetry (39) for $x, y \in B$.

Indeed, $W_\lambda(x,y)$ is a solution of the Helmholtz equation separately as a function of x and of y. Let us introduce a family of solutions $\{w_n(x)\}_{n=0}^{\infty}$ of (40) in B, such that the members of this family form an orthonormal basis for all solutions of the latter equation in B. For example, the spherical waves, i.e., the products of spherical harmonics and Bessel functions, can serve as such a basis.

Then $W_\lambda(x, y)$ can be expanded n the following series:

$$W_\lambda(x, y) = \sum_{n=0}^{\infty} \sum_{m=0}^{\infty} b_{n,m} w_m(y) w_n(x). \tag{41}$$

Since $W_\lambda(x, y)$ is a solution to the Helmholtz equation in $\partial B \times \partial B$, coefficients $b_{n,m}$ are completely determined by the boundary values of W_λ. Since the boundary values are symmetric, the coefficients are symmetric, i.e., $b_{n,m} = b_{m,n}$ which by (41) immediately implies $W_\lambda(x, y) = W_\lambda(y, x)$ for all pairs $(x, y) \in B \times B$.

This consideration extends to infinite cylinders and planes. This explains why the "universal backprojection formula" (30) is valid also for infinite cylinders and planes [98]. Since the sort of symmetry used is shared only by these three smooth surfaces, we believe it is unlikely that a closed-form formula could exist for any other smooth acquisition surface. However, an exact formula has recently been obtained by L. Kunyansky for the case when observation surface S is a surface of a cube (unpublished).

Algebraic Iterative Algorithms

Iterative algebraic techniques are among the favorite tomographic methods of reconstruction and have been used in CT for quite a while [63]. They amount to discretizing the equation relating the measured data with the unknown source, followed by iterative solution of the resulting linear system. Iterative algebraic reconstruction algorithms frequently produce better images than those obtained by other methods. However, they are notoriously slow. In TAT, they have been used successfully for reconstructions with partial data [14, 15, 76], see section "Partial (Incomplete) Data."

Parametrix Approaches

Some of the earlier non-iterative reconstruction techniques [53] were of approximate nature. For example, by approximating the integration spheres by their tangent planes at the point of reconstruction and by applying one of the known inversion formulas for the classical Radon transform, one can reconstruct an approximation to the image. Due to the evenness symmetry in the classical Radon projections (see section "Range Conditions"), the normals to the integration planes need only fill a half of a unit sphere, in order to make possible the reconstruction from an open measurement surface. A more sophisticated approach is represented by the so-called straightening methods [81, 82] based on the approximate reconstruction of the classical Radon projections from the values of the spherical mean transform $\mathcal{M}f$ of the function $f(x)$ in question. These methods yield not a true inversion, but rather what is called in microlocal analysis a **parametrix**. Application of a parametrix reproduces the function f with an additional, smoother term. In other words, the locations (and often the sizes) of jumps across sharp material interfaces, as well as the whole wave front set $WF(f)$, are reconstructed correctly, while the accuracy of the lower spatial frequencies cannot be guaranteed. (Sometimes, the

reconstructed function has a more general form Af, where A is an elliptic pseudo-differential operator [46, 84] of order zero. In this case, the sizes of the jumps across the interfaces might be altered.) Unlike the approximations resulting from the discretization of the exact inversion formulas (in the situations when such formulas are known), the parametrix approximations do not converge, when the discretization of the data is refined and the noise is eliminated. Parametrix reconstructions can be either accepted as approximate images or used as starting points for iterative algorithms.

These methods are closely related to the general scheme proposed in [20] for the inversion of the generalized Radon transform with integration over curved manifolds. It reduces the problem to a Fredholm integral equation of the second kind, which is well suited for numerical solution. Such an approach amounts to using a parametrix method as an efficient preconditioner for an iterative solver; the convergence of such iterations is much faster than that of algebraic iterative methods.

Numerical Implementation and Computational Examples

By discretizing exact formulas presented above, one can easily develop accurate and efficient reconstruction algorithms. The $3D$ case is especially simple: computation of derivatives in the formulas (26)–(30) and (38) can be easily done, for instance, by using finite differences; it is followed by the backprojection (described by the integral over ∂B), which requires prescribing quadrature weights for quadrature nodes that coincide with the positions of the detectors. The backprojection step is stable; the differentiation is a mildly unstable operation. The sensitivity to noise in measurements across the formulas presented above seems to be roughly the same. It is very similar to that of the widely used FBP algorithm of classical X-ray tomography [63]. In $2D$, the implementation is just a little bit harder: the filtration step in formulas (34)–(36) can be reduced to computing two Hilbert transforms (see [54]), which, in turn, can be easily done in the frequency domain.

The number of floating point operations (flops) required by such algorithms is determined by the slower backprojection step. In $3D$, if the number of detectors is m^2 and the size of the reconstruction grid is $m \times m \times m$, the backprojection step (and the whole algorithm) will require $O(m^5)$ flops. In practical terms, this amounts to several hours of computations on a single processor computer for a grid of size $129 \times 129 \times 129$.

In $2D$, the operation count is just $O(m^3)$. As it is discussed in section "Variations on the Theme: Planar, Linear, and Circular Integrating Detectors," the $2D$ problem needs to be solved, when integrating line detectors are used. In this situation, the $2D$ problem needs to be solved m times in order to reconstruct the image, which raises the total operation count to $O(m^4)$ flops.

Figure 10 shows three examples of simulated reconstruction using formula (34). The phantom we use (Fig. 10a) is a linear combination of several characteristic functions of disks and ellipses. Figure 10b illustrates the image reconstruction within the unit circle from 257 equi-spaced projections each containing 129 spherical integrals. The detectors were placed on the concentric circle of radius 1.05.

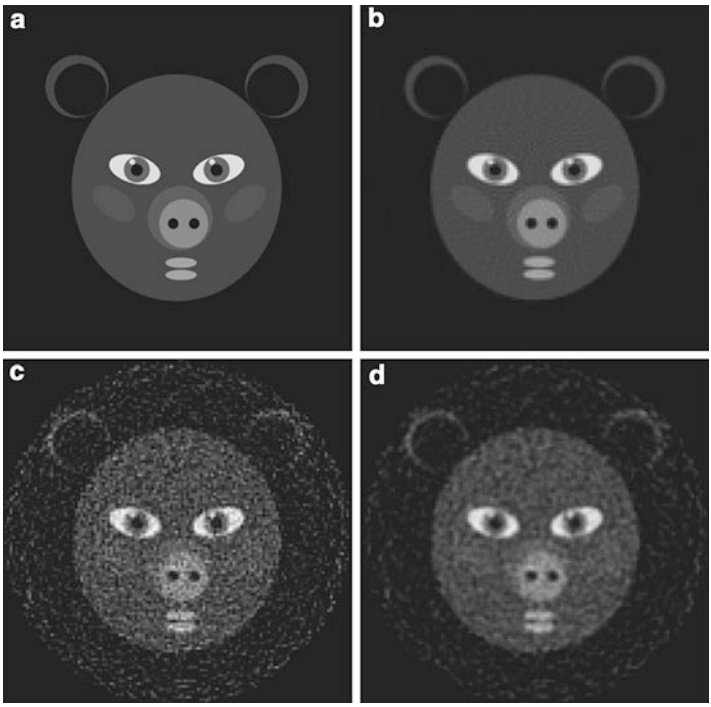


Fig. 10 Example of a reconstruction using formula (34): (**a**) phantom; (**b**) reconstruction from accurate data; (**c**) reconstruction from the data contaminated with 15 % noise; (**d**) reconstruction from the noisy data with additional smoothing

The image shown in Fig. 10c corresponds to the reconstruction from the simulated noisy data that were obtained by adding to projections values of a random variable scaled so that the L^2 intensity of the noise was 15 % of the intensity of the signal. Finally, Fig. 10d shows how the application of a smoothing filter (in the frequency domain) suppresses the noise; it also somewhat blurs the edges in the image.

Variable Speed of Sound

The reconstruction formulas and algorithms described in the previous section work under the assumption that the speed of sound within the region of interest is constant (or at least close to a constant). This assumption, however, is not always realistic, e.g., if the region of interest contains both soft tissues and bones, the speed of sound will vary significantly. Experiments with numerical and physical phantoms show [48, 50] that if acoustic inhomogeneities are not taken into account, the reconstructed image might be severely distorted. Not only the numerical values could be reconstructed incorrectly, but so would the material interface locations and discontinuity magnitudes.

Below we review some of the reconstruction methods that work in acoustically inhomogeneous media. We will assume that the speed of sound $c(x)$ is known,

smooth, positive, constant for large x, and non-trapping. In practice, a transmission ultrasound scan can be used to reconstruct $c(x)$ prior to thermoacoustic reconstruction, as it is done in [50].

Time Reversal

Let us assume temporarily that the speed of sound c is constant and the spatial dimension is odd. Then Huygens' principle guarantees that the sound wave will leave the region of interest Ω in time $T = c/(\text{diam}\,\Omega)$, so that $p(x,t) = 0$ for all $x \in \Omega$ and $t \geq T$. Now one can solve the wave equation back in time from $t = T$ to $t = 0$ in the domain $\Omega \times [T, 0]$, with zero initial conditions at T and boundary conditions on S provided by the data g collected by the detectors. Then the value of the solution at $t = 0$ will coincide with the initial condition $f(x)$ that one seeks to reconstruct. Such a solution of the wave equation is easily obtained numerically by finite difference techniques [42, 48]. The required number of floating point operations is actually lower than that of methods based on discretized inversion formulas ($\mathcal{O}(m^4)$ for time reversal on a grid $m \times m \times m$ in $3D$ versus $\mathcal{O}(m^5)$ for inversion formulas), which makes this method quite competitive even in the case of constant speed of sound.

Most importantly, however, the method is also applicable if the speed of sound $c(x)$ is variable and/or the spatial dimension is even. In these cases, the Huygens' principle does not hold, and thus the solution to the direct problem will not vanish within $\partial\Omega$ in finite time. However, the solution inside Ω will decay with time. Under the non-trapping condition, as it is shown in (11) (see [32,90,91]), the time decay is exponential in odd dimensions, but only algebraic in even dimensions. Although, in order to obtain theoretically exact reconstruction, one would have to start the time reversal at $T = \infty$, and numerical experiments (e.g., [48]) and theoretical estimates [47] show that in practice it is sufficient to start at the values of T when the signal becomes small enough and to approximate the unknown value of $p(x,T)$ by zero (a more sophisticated cutoff is used in [86]). This works [42,48] even in $2D$ (where decay is the slowest) and in inhomogeneous media. However, when trapping occurs, the "invisible" parts blur away (see section "Incomplete Data" for the discussion).

Eigenfunction Expansions

An "inversion formula" that reconstructs the initial value $f(x)$ of the solution of the wave equation from values on the measuring surface S can be easily obtained using time reversal and Duhamel's principle [3]. Consider in Ω the operator $A = -c^2(x)\Delta$ with zero Dirichlet conditions on the boundary $S = \partial\Omega$. This operator is self-adjoint, if considered in the weighted space $L^2(\Omega; c^{-2}(x))$. Let us denote by E the operator of harmonic extension, which transforms a function ϕ on S to a harmonic function on Ω that coincides with ϕ on S. Then f can be reconstructed [3] from the data g in (3) by the following formula:

$$f(x) = (Eg|_{t=0}) - \int_0^\infty A^{-\frac{1}{2}} \sin\left(\tau A^{\frac{1}{2}}\right) E(g_{tt})(x,\tau)\, d\tau, \tag{42}$$

which is valid under the non-trapping condition on $c(x)$. However, due to the involvement of functions of the operator A, it is not clear how useful this formula can be.

One natural way to try to implement numerically the formula (42) is to use the eigenfunction expansion of the operator A in Ω (assuming that such expansion is known). This quickly leads to the following procedure [3]. The function $f(x)$ can be reconstructed inside Ω from the data g in (3), as the following $L^2(B)$-convergent series:

$$f(x) = \sum_k f_k \psi_k(x), \tag{43}$$

where the Fourier coefficients f_k can be recovered from the data using one of the following formulas:

$$\begin{gathered} f_k = \lambda_k^{-2} g_k(0) - \lambda_k^{-3} \int_0^\infty \sin(\lambda_k t) g_k''(t) dt, \\ f_k = \lambda_k^{-2} g_k(0) + \lambda_k^{-2} \int_0^\infty \cos(\lambda_k t) g_k'(t) dt, \text{ or} \\ f_k = -\lambda_k^{-1} \int_0^\infty \sin(\lambda_k t) g_k(t) dt = -\lambda_k^{-1} \int_0^\infty \int_S \sin(\lambda_k t) g(x,t) \overline{\frac{\partial \psi_k}{\partial n}(x)} dx dt, \end{gathered} \tag{44}$$

where

$$g_k(t) = \int_S g(x,t) \overline{\frac{\partial \psi_k}{\partial n}(x)} dx.$$

One notices that this is a generalization of the expansion method of [58] discussed in section "Eigenfunction Expansions for a General Geometry" to the case of a variable speed of sound. Unlike the algorithm of [58], this method does not require the knowledge of the whole space Green's function for A (which is in this case unknown). However, computation of a large set of eigenfunctions and eigenvalues followed by the summation of the series (43) at the nodes of the computational grid may prove to be too time-consuming.

It is worthwhile to mention again that the non-trapping condition is crucial for the stability of any TAT reconstruction method in acoustically inhomogeneous media. As it was discussed in section "Discussion of the Visibility Condition," trapping can significantly reduce the quality of reconstruction. It is, however, most probable that trapping does not occur much in biological objects.

Partial (Incomplete) Data

Reconstruction formulas and algorithms of the previous sections work under the assumption that the acoustic signal is measured by detectors covering a closed

surface S that surrounds completely the object of interest. However, in many practical applications of TAT, detectors can be placed only on a certain part of the surrounding surface. Such is the case, e.g., when TAT is used for breast screening – one of the most promising applications of this modality. Thus, one needs methods and algorithms capable of accurate reconstruction of images from partial (incomplete) data, i.e., from the measurements made on open surfaces (or open curves in $2D$).

Most exact inversion formulas and methods discussed above are based (explicitly or implicitly) on some sort of the Green's formula, Helmholtz representation, or eigenfunction decomposition for closed surfaces, and thus they cannot be extended to the case of partial data. The methods that do work in this situation rely on approximation techniques, as discussed below.

Constant Speed of Sound

Even the case of an acoustically homogeneous medium is quite challenging when reconstruction needs to be done from partial data (i.e., when the acquisition surface S is not closed). As it was discussed in section "Incomplete Data," if the detectors located around the object in such a way that the "visibility" condition is not satisfied, accurate reconstruction is impossible: the "invisible" interfaces will be smoothed out in the reconstructed image. On the other hand, if the visibility condition is satisfied, the reconstruction is only mildly unstable (similarly to the inversion of the classic Radon transform) [71, 86]. If, in addition, the uniqueness of reconstruction from partial data is guaranteed (which is usually the case, see section "Uniqueness of Reconstruction"), one can hope to be able to develop an algorithm that would reconstruct quality images.

Special cases of open acquisition surfaces are a plane or an infinite cylinder, for which exact inversion formulas are known (see, e.g., [16,34,41,96] for the plane and [101] for a cylinder). Of course, the plane or a cylinder would have to be truncated in any practical measurements. The resulting acquisition geometry will not satisfy the visibility condition, and material interfaces whose normals do not intersect the acquisition surface will be blurred.

Iterative algebraic techniques (see the corresponding paragraph in section "Constant Speed of Sound") were among the first methods successfully used for reconstruction from surfaces only partially surrounding the object (e.g., [14,15,76]). As it is mentioned in section "Constant Speed of Sound," such methods are very slow. For example, reconstructions in [14] required the use of a cluster of computers and took 100 iterations to converge.

Parametrix-type reconstructions in the partial data case were proposed in [17]. A couple of different parametrix-type algorithms were proposed in [72, 74]. They are based on applying one of the exact inversion formulas for full circular acquisition to the available partial data, with zero-filled missing data, and some correction factors. Namely, since the missing data is replaced by zeros, each line passing through a node of the reconstruction grid will be tangent either to one or to two circles of integration. Therefore, some directions during the backprojection step will be

represented twice and some only once. This, in turn, will cause some interfaces to appear twice stronger than they should be. The use of weight factors was proposed in [72, 74] in order to partially compensate for this distortion. In particular, in [72] smooth weight factors (depending on a reconstruction point) are assigned to each detector in such a way that the total weight for each direction is exactly one. This method is not exact; the error is described by a certain smoothing operator. However, the singularities (or jumps) in the image will be reconstructed correctly. As shown by numerical examples in [72], such a correction visually significantly improves the reconstruction. Moreover, iterative refinement is proposed in [72, 74] to further improve the image, and it is shown to work well in numerical experiments.

Returning to non-iterative techniques, one should mention an interesting attempt made in [78, 79] to generate the missing data using the moment range conditions for $\mathcal{M}$ (see section "Range Conditions"). The resulting algorithm, however, does not seem to recover the values well, although, as expected, it reconstructs all visible singularities.

An accurate $2D$ non-iterative algorithm for reconstruction from data measured on an open curve S was proposed in [59]. It is based on precomputing approximations of plane waves in the region of interest Ω by the single-layer potentials of the form

$$\int_S Z(\lambda|y-x|)\rho(y)dl(y),$$

where $\rho(y)$ is the density of the potential, which needs to be chosen appropriately, $dl(y)$ is the standard arc length, and $Z(t)$ is either the Bessel function $J_0(t)$ or the Neumann function $Y_0(t)$. Namely, for a fixed ξ one finds numerically the densities $\rho_{\xi,J}(y)$ and $\rho_{\xi,Y}(y)$ of the potentials

$$W_J(x,\rho_{\xi,J}) = \int_S J_0(\lambda|y-x|)\rho_{\xi,J}(y)dl(y), \tag{45}$$

$$W_Y(x,\rho_{\xi,Y}) = \int_S Y_0(\lambda|y-x|)\rho_{\xi,Y}(y)dl(y), \tag{46}$$

where $\lambda = |\xi|$, such that

$$W_J(x,\rho_{\xi,J}) + W_Y(x,\rho_{\xi,Y}) \approx \exp(-i\xi\cdot x) \text{ for all } x\in\Omega. \tag{47}$$

Obtaining such approximations is not trivial. One can show that exact equality in (47) cannot be achieved, due to different behavior at infinity of the plane wave and the approximating single-layer potentials. However, as shown by numerical examples in [59], if each point in Ω is "visible" from S, very accurate *approximations* can be obtained, while keeping the densities $\rho_{\xi,J}$ and $\rho_{\xi,Y}$ under certain control.

Once the densities $\rho_{\xi,J}$ and $\rho_{\xi,Y}$ have been found for all ξ, function $f(x)$ can be easily reconstructed. Indeed, for the Fourier transform $\hat{f}(\xi)$ of $f(x)$

$$\hat{f}(\xi) = \frac{1}{2\pi} \int_{\Omega} f(x) \exp(-i\xi \cdot x) dx,$$

one obtains, using (47)

$$\begin{aligned} \hat{f}(\xi) &\approx \frac{1}{2\pi} \int_{\Omega} f(x) \left[W_J(x, \rho_{\xi,J}) + W_Y(x, \rho_{\xi,Y}) \right] dx \\ &= \frac{1}{2\pi} \int_S \left[\int_{\Omega} f(x) J_0(\lambda |y - x|) dx \right] \rho_{\xi,J}(y) dl(y) \\ &\quad + \frac{1}{2\pi} \int_S \left[\int_{\Omega} f(x) Y_0(\lambda |y - x|) dx \right] \rho_{\xi,Y}(y) dl(y), \end{aligned} \tag{48}$$

where the inner integrals are computed from the data g

$$\int_{\Omega} f(x) J_0(\lambda |y - x|) dx = \int_{R^+} g(y, r) J_0(\lambda r) dr, \tag{49}$$

$$\int_{\Omega} f(x) Y_0(\lambda |y - x|) dx = \int_{R^+} g(y, r) Y_0(\lambda r) dr. \tag{50}$$

Formula (48), in combination with (49) and (50), yields values of $\hat{f}(\xi)$ for arbitrary ξ. Now $f(x)$ can be recovered by numerically inverting the Fourier transform or by a reduction to an FBP inversion [63] of the regular Radon transform.

The most computationally expensive part of the algorithm, which is computing the densities $\rho_{\xi,J}$ and $\rho_{\xi,Y}$, needs to be done only once for a given acquisition surface. Thus, for a scanner with a fixed S, the resulting densities can be precomputed once and for all. The actual reconstruction part then becomes extremely fast.

Examples of reconstructions from incomplete data using this technique of [59] are shown in Fig. 11. The images were reconstructed within the unit square $[-1, 1] \times [-1, 1]$, while the detectors were placed on the part of the concentric circle of radius 1.3 lying to the left of line $x_1 = 1$. We used the same phantom as in Fig. 10a; the reconstruction from the data with added 15 % noise is shown in Fig. 11b; Fig. 11c demonstrates the results of applying additional smoothing filter to reduce the effects of noise in the data.

Variable Speed of Sound

The problem of numerical reconstruction in TAT from the data measured on open surfaces in the presence of a known variable speed of sound currently remains largely open. One of the difficulties was discussed in section "Incomplete Data": even if the speed of sound $c(x)$ is non-trapping, it can happen that some of the characteristics escape from the region of interest to infinity without intersecting the open measuring surface. Then stable reconstruction of the corresponding interfaces will become impossible. It should be possible, however, to develop stable

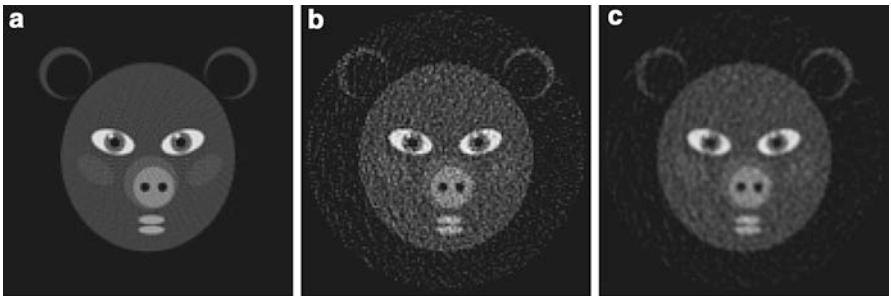


Fig. 11 Examples of reconstruction from incomplete data using the technique of [59]. Detectors are located on the part of circular arc of radius 1.3 lying left of the line $x_1 = 1$. (**a**) Reconstruction from accurate data, (**b**) reconstruction from the data with added 15 % noise, (**c**) reconstruction from noisy data with additional smoothing filter

reconstruction algorithms in the case when the whole object of interest is located in the visible zone.

The generalization of the method of [59] to the case of variable speed of sound is so far problematic, since this algorithm is based on the knowledge of the open space Green's function for the Helmholtz equation. In the case of a nonconstant $c(x)$, this Green's function is position dependent, and its numerical computation is likely to be prohibitively time-consuming.

A promising approach to this problem, currently under development, is to use time reversal with the missing data replaced by zeros or maybe by a more clever extension (e.g., using the range conditions, as in [78, 79]). This would produce an initial approximation to $f(x)$, which one can try to refine by fixed-point iterations; however, the pertinent questions concerning such an algorithm remain open.

An interesting technique of using a reverberant cavity enclosing the target to compensate for the missing data is described in [28].

5 Final Remarks and Open Problems

We list here some unresolved issues of mathematics of TAT/PAT, as well as some developments that were not addressed in the main text.

1. The issue of uniqueness acquisition sets S (i.e., such that transducers distributed along S provide sufficient information for TAT reconstruction) can be considered to be resolved, for most practical purposes. However, there remain significant unresolved theoretical questions. One of them consists of proving an analogue of Theorem 5 for non-compactly supported functions with a sufficiently fast (e.g., super-exponential) decay at infinity. The original (and the only known) proof of this theorem uses microlocal techniques [7, 85] that significantly rely upon the compactness of support. However, one hopes that the condition of a fast

decay should suffice for this result. In particular, there is no proven analogue of Theorem 3 for non-closed sets S (unless S is an open part of a closed analytic surface).

Techniques developed in [36] (see also [8] for their further use in TAT) might provide the right approach.

This also relates to still unresolved situation in dimensions 3 and higher. Namely, one would like to prove Conjecture 1.

2. Concerning the inversion methods, one notices that closed-form formulas are known only for spherical, cylindrical, and planar acquisition surfaces. The question arises whether closed-form inversion formulas could be found for any other smooth closed surface? It is the belief of the authors that the answer to this question is negative.

   Another feature of the known closed-form formulas that was mentioned before is that they do not work correctly if the support of the sought function $f(x)$ lies partially outside the acquisition surface. Time reversal and eigenfunction expansion methods do not suffer from this deficiency. The question arises whether one could find closed-form formulas that reconstruct the function inside S correctly, in spite of it having part of its support outside. Again, the authors believe that the answer is negative.
3. The complete range description of the forward operator $\mathcal{W}$ in even dimensions is still not known. It is also not clear whether one can obtain complete range descriptions for nonspherical observation sets S or for a variable sound speed. The moment and orthogonality conditions do hold in the case of a constant speed and arbitrary closed surface, but they do not provide a complete description of the range. For acoustically inhomogeneous media, an analogue of orthogonality conditions exists, but it also does not describe the range completely.
4. The problem of unique determination of the speed of sound from TAT data is largely open.
5. As it was explained in the text, knowing full Cauchy data of the pressure p (i.e., its value and the value of its the normal derivative) on the observation surface S leads to unique determination and simple reconstruction of f. However, the normal derivative is not measured by transducers and thus needs to be either found mathematically or measured in a different experiment. Thus, feasibility of techniques [12,25] relying on full Cauchy data requires further mathematical and experimental study.
6. In the standard X-ray CT, as well as in SPECT, the **local tomography** technique [33, 56] is often very useful. It allows one to emphasize in a stable way singularities (e.g., tissue interfaces) of the reconstruction, even in the case of incomplete data (in the latter case, the invisible parts will be lost). An analogue of local tomography can be easily implemented in TAT, for instance, by introducing an additional high-pass filter in the FBP-type formulas.
7. The mathematical analysis of TAT presented in the text did not take into account the issue of modeling and compensating for the **acoustic attenuation**. This subject is addressed in [22, 52, 62, 80, 83], but probably cannot be considered completely resolved.

8. **Quantitative PAT**: This chapter, as well as most other papers devoted to TAT/PAT, is centered on finding the initial pressure $f(x)$. This pressure, which is proportional to the initial energy deposition, is related to the optical parameters (attenuation and scattering coefficients) of the tissue. The nontrivial issue of recovering these parameters, after the initial pressure $f(x)$ is found, is addressed in the recent works [18, 29, 30].
9. The TAT technique discussed in the chapter uses active interrogation of the medium. There is a discussion in the literature of a **passive version of TAT**, where no irradiation of the target is involved [77].

Acknowledgments The work of both authors was partially supported by the NSF DMS grant 0908208. The first author was also supported by the NSF DMS grant 0604778 and by the KAUST grant KUS-CI-016-04 through the IAMCS. The work of the second author was partially supported by the DOE grant DE-FG02-03ER25577. The authors express their gratitude to NSF, DOE, KAUST, and IAMCS for the support.

Cross-References

▸ Linear Inverse Problems
▸ Microlocal Analysis in Tomography
▸ Photoacoustic and Thermoacoustic Tomography: Image Formation Principles
▸ Tomography

References

1. Agranovsky, M., Berenstein, C., Kuchment, P.: Approximation by spherical waves in L^p-spaces. J. Geom. Anal. **6**(3), 365–383 (1996)
2. Agranovsky, M., Finch, D., Kuchment, P.: Range conditions for a spherical mean transform. Inverse Probl. Imaging **3**(3), 373–38 (2009)
3. Agranovsky, M., Kuchment, P.: Uniqueness of reconstruction and an inversion procedure for thermoacoustic and photoacoustic tomography with variable sound speed. Inverse Probl. **23**, 2089–2102 (2007)
4. Agranovsky, M., Kuchment, P., Kunyansky, L.: On reconstruction formulas and algorithms for the thermoacoustic and photoacoustic tomography, chapter 8. In: Wang, L.H. (ed.) Photoacoustic Imaging and Spectroscopy, pp. 89–101. CRC, Boca Raton (2009)
5. Agranovsky, M., Kuchment, P., Quinto, E.T.: Range descriptions for the spherical mean Radon transform. J. Funct. Anal. **248**, 344–386 (2007)
6. Agranovsky, M., Nguyen, L.: Range conditions for a spherical mean transform and global extension of solutions of Darboux equation. J. d'Analyse Math. (2009). Preprint arXiv:0904.4225 (to appear)
7. Agranovsky, M., Quinto, E.T.: Injectivity sets for the Radon transform over circles and complete systems of radial functions. J. Funct. Anal. **139**, 383–414 (1996)
8. Ambartsoumian, G., Kuchment, P.: On the injectivity of the circular Radon transform. Inverse Probl. **21**, 473–485 (2005)
9. Ambartsoumian, G., Kuchment, P.: A range description for the planar circular Radon transform. SIAM J. Math. Anal. **38**(2), 681–692 (2006)

10. Ammari, H.: An Introduction to Mathematics of Emerging Biomedical Imaging. Springer, Berlin (2008)
11. Ammari, H., Bonnetier, E., Capdebosq, Y., Tanter, M., Fink, M.: Electrical impedance tomography by elastic deformation. SIAM J. Appl. Math. **68**(6), 1557–1573 (2008)
12. Ammari, H., Bossy, E., Jugnon, V., Kang, H.: Quantitative photo-acoustic imaging of small absorbers. SIAM Rev. (to appear)
13. Anastasio, M.A., Zhang, J., Modgil, D., Rivière, P.J.: Application of inverse source concepts to photoacoustic tomography. Inverse Probl. **23**, S21–S35 (2007)
14. Anastasio, M., Zhang, J., Pan, X., Zou, Y., Ku, G., Wang, L.V.: Half-time image reconstruction in thermoacoustic tomography. IEEE Trans. Med. Imaging **24**, 199–210 (2005)
15. Anastasio, M.A., Zhang, J., Sidky, E.Y., Zou, Z., Dan, X., Pan, X.: Feasibility of half-data image reconstruction in 3-D reflectivity tomography with a spherical aperture. IEEE Trans Med. Imaging **24**(9), 1100–1112 (2005)
16. Andersson, L.-E.: On the determination of a function from spherical averages. SIAM J. Math. Anal. **19**(1), 214–232 (1988)
17. Andreev, V., Popov, D., et al.: Image reconstruction in 3D optoacoustic tomography system with hemispherical transducer array. Proc. SPIE **4618**, 137–145 (2002)
18. Bal, G., Jollivet, A., Jugnon, V.: Inverse transport theory of photoacoustics. Inverse Probl. **26**, 025011 (2010). doi:10.1088/0266-5611/26/2/025011
19. Bell, A.G.: On the production and reproduction of sound by light. Am. J. Sci. **20**, 305–324 (1880)
20. Beylkin, G.: The inversion problem and applications of the generalized Radon transform. Commun. Pure Appl. Math. **37**, 579–599 (1984)
21. Bowen, T.: Radiation-induced thermoacoustic soft tissue imaging. Proc. IEEE Ultrason. Symp. **2**, 817–822 (1981)
22. Burgholzer, P., Grün, H., Haltmeier, M., Nuster, R., Paltauf, G.: Compensation of acoustic attenuation for high-resolution photoacoustic imaging with line detectors using time reversal. In: Proceedings of the SPIE Number 6437–75 Photonics West, BIOS 2007, San Jose (2007)
23. Burgholzer, P., Hofer, C., Matt, G.J., Paltauf, G., Haltmeier, M., Scherzer, O.: Thermoacoustic tomography using a fiber-based Fabry–Perot interferometer as an integrating line detector. Proc. SPIE **6086**, 434–442 (2006)
24. Burgholzer, P., Hofer, C., Paltauf, G., Haltmeier, M., Scherzer, O.: Thermoacoustic tomography with integrating area and line detectors. IEEE Trans. Ultrason. Ferroelectr. Freq. Control **52**(9), 1577–1583 (2005)
25. Clason, C., Klibanov, M.: The quasi-reversibility method in thermoacoustic tomography in a heterogeneous medium. SIAM J. Sci. Comput. **30**, 1–23 (2007)
26. Colton, D., Paivarinta, L., Sylvester, J.: The interior transmission problem. Inverse Probl. **1**(1), 13–28 (2007)
27. Courant, R., Hilbert, D.: Methods of Mathematical Physics. Partial Differential Equations, vol. II. Interscience, New York (1962)
28. Cox, B.T., Arridge, S.R., Beard, P.C.: Photoacoustic tomography with a limited aperture planar sensor and a reverberant cavity. Inverse Probl. **23**, S95–S112 (2007)
29. Cox, B.T., Arridge, S.R., Beard, P.C.: Estimating chromophore distributions from multiwavelength photoacoustic images. J. Opt. Soc. Am. A **26**, 443–455 (2009)
30. Cox, B.T., Laufer, J.G., Beard, P.C.: The challenges for quantitative photoacoustic imaging. Proc. SPIE **7177**, 717713 (2009)
31. Diebold, G.J., Sun, T., Khan, M.I.: Photoacoustic monopole radiation in one, two, and three dimensions. Phys. Rev. Lett. **67**(24), 3384–3387 (1991)
32. Egorov, Yu.V., Shubin, M.A.: Partial Differential Equations I. Encyclopaedia of Mathematical Sciences, vol. 30, pp. 1–259. Springer, Berlin (1992)
33. Faridani, A., Ritman, E.L., Smith, K.T.: Local tomography. SIAM J. Appl. Math. **52**(4), 459–484 (1992)
34. Fawcett, J.A.: Inversion of n-dimensional spherical averages. SIAM J. Appl. Math. **45**(2), 336–341 (1985)

35. Finch, D., Haltmeier, M., Rakesh: Inversion of spherical means and the wave equation in even dimensions. SIAM J. Appl. Math. **68**(2), 392–412 (2007)
36. Finch, D., Patch, S., Rakesh: Determining a function from its mean values over a family of spheres. SIAM J. Math. Anal. **35**(5), 1213–1240 (2004)
37. Finch, D., Rakesh: Range of the spherical mean value operator for functions supported in a ball. Inverse Probl. **22**, 923–938 (2006)
38. Finch, D., Rakesh: Recovering a function from its spherical mean values in two and three dimensions. In: Wang, L. (ed.) Photoacoustic Imaging and Spectroscopy, pp. 77–88. CRC, Boca Raton (2009)
39. Finch, D., Rakesh: The spherical mean value operator with centers on a sphere. Inverse Probl. **23**(6), S37–S50 (2007)
40. Gebauer, B., Scherzer, O.: Impedance-acoustic tomography. SIAM J. Appl. Math. **69**(2), 565–576 (2009)
41. Gelfand, I., Gindikin, S., Graev, M.: Selected Topics in Integral Geometry. Translations of Mathematical Monographs, vol. 220. American Mathematical Society, Providence (2003)
42. Grün, H., Haltmeier, M., Paltauf, G., Burgholzer, P.: Photoacoustic tomography using a fiber based Fabry-Perot interferometer as an integrating line detector and image reconstruction by model-based time reversal method. Proc. SPIE **6631**, 663107 (2007)
43. Haltmeier, M., Burgholzer, P., Paltauf, G., Scherzer, O.: Thermoacoustic computed tomography with large planar receivers. Inverse Probl. **20**, 1663–1673 (2004)
44. Haltmeier, M., Scherzer, O., Burgholzer, P., Nuster, R., Paltauf, G.: Thermoacoustic tomography and the circular Radon transform: exact inversion formula. Math. Models Methods Appl. Sci. **17**(4), 635–655 (2007)
45. Helgason, S.: The Radon Transform. Birkhäuser, Basel (1980)
46. Hörmander, L.: The Analysis of Linear Partial Differential Operators, vols. 1 and 2. Springer, New York (1983)
47. Hristova, Y.: Time reversal in thermoacoustic tomography: error estimate. Inverse Probl. **25**, 1–14 (2009)
48. Hristova, Y., Kuchment, P., Nguyen, L.: On reconstruction and time reversal in thermoacoustic tomography in homogeneous and non-homogeneous acoustic media. Inverse Probl. **24**, 055006 (2008)
49. Isakov, V.: Inverse Problems for Partial Differential Equations, 2nd edn. Springer, Berlin (2005)
50. Jin, X., Wang, L.V.: Thermoacoustic tomography with correction for acoustic speed variations. Phys. Med. Biol. **51**, 6437–6448 (2006)
51. John, F.: Plane Waves and Spherical Means Applied to Partial Differential Equations. Dover, New York (1971)
52. Kowar, R., Scherzer, O., Bonnefond, X.: Causality analysis of frequency dependent wave attenuation. Preprint arXiv:0906.4678
53. Kruger, R.A., Liu, P., Fang, Y.R., Appledorn, C.R.: Photoacoustic ultrasound (PAUS)reconstruction tomography. Med. Phys. **22**, 1605–1609 (1995)
54. Kuchment, P., Kunyansky, L.: Mathematics of thermoacoustic tomography. Eur. J. Appl. Math. **19**(02), 191–224 (2008)
55. Kuchment, P., Kunyansky, L.: Synthetic focusing in ultrasound modulated tomography. Inverse Probl. Imaging (to appear)
56. Kuchment, P., Lancaster, K., Mogilevskaya, L.: On local tomography. Inverse Probl. **11**, 571–589 (1995)
57. Kunyansky, L.: Explicit inversion formulae for the spherical mean Radon transform. Inverse probl. **23**, 737–783 (2007)
58. Kunyansky, L.: A series solution and a fast algorithm for the inversion of the spherical mean Radon transform. Inverse Probl. **23**, S11–S20 (2007)
59. Kunyansky, L.: Thermoacoustic tomography with detectors on an open curve: an efficient reconstruction algorithm. Inverse Probl. **24**(5), 055021 (2008)

60. Lin, V., Pinkus, A.: Approximation of multivariate functions. In: Dikshit, H.P., Micchelli, C.A. (eds.) Advances in Computational Mathematics, pp. 1–9. World Scientific, Singapore (1994)
61. Louis, A.K., Quinto, E.T.: Local tomographic methods in Sonar. In: Surveys on Solution Methods for Inverse Problems, pp. 147–154. Springer, Vienna (2000)
62. Maslov, K., Zhang, H.F., Wang, L.V.: Effects of wavelength-dependent fluence attenuation on the noninvasive photoacoustic imaging of hemoglobin oxygen saturation in subcutaneous vasculature in vivo. Inverse Probl. **23**, S113–S122 (2007)
63. Natterer, F.: The Mathematics of Computerized Tomography. Wiley, New York (1986)
64. Nguyen, L.: A family of inversion formulas in thermoacoustic tomography. Inverse Probl. Imaging **3**(4), 649–675 (2009)
65. Nguyen, L.V.: On singularities and instability of reconstruction in thermoacoustic tomography. Preprint arXiv:0911.5521v1
66. Norton, S.J.: Reconstruction of a two-dimensional reflecting medium over a circular domain: exact solution. J. Acoust. Soc. Am. **67**, 1266–1273 (1980)
67. Norton, S.J., Linzer, M.: Ultrasonic reflectivity imaging in three dimensions: exact inverse scattering solutions for plane, cylindrical, and spherical apertures. IEEE Trans. Biomed. Eng. **28**, 200–202 (1981)
68. Olafsson, G., Quinto, E.T. (eds.): The Radon Transform, Inverse Problems, and Tomography. American Mathematical Society Short Course, Atlanta, 3–4 Jan 2005. Proceedings of Symposia in Applied Mathematics, vol. 63. American Mathematical Society, Providence (2006)
69. Oraevsky, A.A., Jacques, S.L., Esenaliev, R.O., Tittel, F.K.: Laser-based photoacoustic imaging in biological tissues. Proc. SPIE **2134A**, 122–128 (1994)
70. Palamodov, V.P.: Reconstructive Integral Geometry. Birkhäuser, Basel (2004)
71. Palamodov, V.: Remarks on the general Funk–Radon transform and thermoacoustic tomography (2007). Preprint arxiv: math.AP/0701204
72. Paltauf, G., Nuster, R., Burgholzer, P.: Weight factors for limited angle photoacoustic tomography. Phys. Med. Biol. **54**, 3303–3314 (2009)
73. Paltauf, G., Nuster, R., Haltmeier, M., Burgholzer, P.: Thermoacoustic computed tomography using a Mach–Zehnder interferometer as acoustic line detector. Appl. Opt. **46**(16), 3352–3358 (2007)
74. Paltauf, G., Nuster, R., Haltmeier, M., Burgholzer, P.: Experimental evaluation of reconstruction algorithms for limited view photoacoustic tomography with line detectors. Inverse Probl. **23**, S81–S94 (2007)
75. Paltauf, G., Nuster, R., Burgholzer, P.: Characterization of integrating ultrasound detectors for photoacoustic tomography. J. Appl. Phys. **105**, 102026 (2009)
76. Paltauf, G., Viator, J.A., Prahl, S.A., Jacques, S.L.: Iterative reconstruction algorithm for optoacoustic imaging J. Acoust. Soc. Am. **112**(4), 1536–1544 (2002)
77. Passechnik, V.I., Anosov, A.A., Bograchev, K.M.: Fundamentals and prospects of passive thermoacoustic tomography. Crit. Rev. Biomed. Eng. **28**(3–4), 603–640 (2000)
78. Patch, S.K.: Thermoacoustic tomography – consistency conditions and the partial scan problem. Phys. Med. Biol. **49**, 1–11 (2004)
79. Patch, S.: (2009) Photoacoustic or thermoacoustic tomography: consistency conditions and the partial scan problem. In: Wang, L. (ed.) Photoacoustic Imaging and Spectroscopy, pp. 103–116. CRC, Boca Raton (2009)
80. Patch, S.K., Haltmeier, M.: Thermoacoustic tomography – ultrasound attenuation artifacts. IEEE Nucl. Sci. Symb. Conf. **4**, 2604–2606 (2006)
81. Popov, D.A., Sushko, D.V.: A parametrix for the problem of optical-acoustic tomography. Dokl. Math. **65**(1), 19–21 (2002)
82. Popov, D.A., Sushko, D.V.: Image restoration in optical-acoustic tomography. Probl. Inf. Transm. **40**(3), 254–278 (2004)
83. La Rivière, P.J., Zhang, J., Anastasio, M.A.: Image reconstruction in optoacoustic tomography for dispersive acoustic media. Opt. Lett. **31**(6), 781–783 (2006)

84. Shubin, M.A.: Pseudodifferential Operators and Spectral Theory. Springer, Berlin (2001)
85. Stefanov, P., Uhlmann, G.: Integral geometry of tensor fields on a class of non-simple Riemannian manifolds. Am. J. Math. **130**(1), 239–268 (2008)
86. Stefanov, P., Uhlmann, G.: Thermoacoustic tomography with variable sound speed. Inverse Probl. **25**, 075011 (2009)
87. Steinhauer, D.: A uniqueness theorem for thermoacoustic tomography in the case of limited boundary data. Preprint arXiv:0902.2838
88. Tam, A.C.: Applications of photoacoustic sensing techniques. Rev. Mod. Phys. **58**(2), 381–431 (1986)
89. Tuchin, V.V. (ed.): Handbook of Optical Biomedical Diagnostics. SPIE, Bellingham (2002)
90. Vainberg, B.: The short-wave asymptotic behavior of the solutions of stationary problems, and the asymptotic behavior as $t \to \infty$ of the solutions of nonstationary problems. Russ. Math. Surv. **30**(2), 1–58 (1975)
91. Vainberg, B.: Asymptotics Methods in the Equations of Mathematical Physics. Gordon & Breach, New York (1982)
92. Vo-Dinh, T. (ed.): Biomedical Photonics Handbook. CRC, Boca Raton (2003)
93. Wang, L. (ed.): Photoacoustic Imaging and Spectroscopy. CRC, Boca Raton (2009)
94. Wang, K., Anastasio, M.A.: Photoacoustic and thermoacoustic tomography: image formation principles. In: Scherzer, O. (ed.) Handbook of Mathematical Methods in Imaging, Chapter 18, pp. 781–815. Springer, New York (2011)
95. Wang, L.V., Wu, H.: Biomedical Optics. Principles and Imaging. Wiley, New York (2007)
96. Xu, Y., Feng, D., Wang, L.-H.V.: Exact frequency-domain reconstruction for thermoacoustic tomography: I planar geometry. IEEE Trans. Med. Imaging **21**, 823–828 (2002)
97. Xu, M., Wang, L.-H.V.: Time-domain reconstruction for thermoacoustic tomography in a spherical geometry. IEEE Trans. Med. Imaging **21**, 814–822 (2002)
98. Xu, M., Wang, L.-H.V.: Universal back-projection algorithm for photoacoustic computed tomography. Phys. Rev. E **71**, 016706 (2005)
99. Xu, Y., Wang, L., Ambartsoumian, G., Kuchment, P.: Reconstructions in limited view thermoacoustic tomography. Med. Phys. **31**(4), 724–733 (2004)
100. Xu, Y., Wang, L., Ambartsoumian, G., Kuchment, P.: Limited view thermoacoustic tomography, Ch. 6. In: Wang, L.H. (ed.) Photoacoustic Imaging and Spectroscopy, pp. 61–73. CRC, Boca Raton (2009)
101. Xu, Y., Xu, M., Wang, L.-H.V.: Exact frequency-domain reconstruction for thermoacoustic tomography: II cylindrical geometry. IEEE Trans. Med. Imaging **21**, 829–833 (2002)
102. Yuan, Z., Zhang, Q., Jiang, H.: Simultaneous reconstruction of acoustic and optical properties of heterogeneous media by quantitative photoacoustic tomography. Opt. Express **14**(15), 6749 (2006)
103. Zangerl, G., Scherzer, O., Haltmeier, M.: Circular integrating detectors in photo and thermoacoustic tomography. Inverse Probl. Sci. Eng. **17**(1), 133–142 (2009)
104. Zhang, J., Anastasio, M.A.: Reconstruction of speed-of-sound and electromagnetic absorption distributions in photoacoustic tomography. Proc. SPIE **6086**, 608619 (2006)

Mathematical Methods of Optical Coherence Tomography

Peter Elbau, Leonidas Mindrinos, and Otmar Scherzer

Contents

1 Introduction 1170
2 Basic Principles of OCT 1171
3 The Direct Scattering Problem 1172
Maxwell's Equations 1172
Initial Conditions 1174
The Measurements 1175
4 Solution of the Direct Problem 1178
Born and Far Field Approximation 1179
The Forward Operator 1181
5 The Inverse Scattering Problem 1185
The Isotropic Case 1187
The Anisotropic Case 1196
6 Conclusion 1202
Cross-References 1202
References 1202

Abstract

In this chapter a general mathematical model of Optical Coherence Tomography (OCT) is presented on the basis of the electromagnetic theory. OCT produces high-resolution images of the inner structure of biological tissues. Images are obtained by measuring the time delay and the intensity of the backscattered light from the sample considering also the coherence properties of light. The

P. Elbau (✉) • L. Mindrinos
Computational Science Center, University of Vienna, Vienna, Austria
e-mail: peter.elbau@univie.ac.at; leonidas.mindrinos@univie.ac.at

O. Scherzer
Computational Science Center, University of Vienna, Vienna, Austria

RICAM, Austrian Academy of Sciences, Linz, Austria
e-mail: otmar.scherzer@univie.ac.at

© Springer Science+Business Media New York 2015

O. Scherzer (ed.), *Handbook of Mathematical Methods in Imaging*,
DOI 10.1007/978-1-4939-0790-8_44

scattering problem is considered for a weakly scattering medium located far enough from the detector. The inverse problem is to reconstruct the susceptibility of the medium given the measurements for different positions of the mirror. Different approaches are addressed depending on the different assumptions made about the optical properties of the sample. This procedure is applied to a full field OCT system and an extension to standard (time and frequency domain) OCT is briefly presented.

1 Introduction

Optical Coherence Tomography (OCT) is a noninvasive imaging technique producing high-resolution images of biological tissues. OCT is based on Low (time) Coherence Interferometry and takes into account the coherence properties of light to image microstructures with resolution in the range of few micrometers. Standard OCT operates using broadband and continuous wave light in the visible and near-infrared spectrum. OCT images are obtained by measuring the time delay and the intensity of backscattered or back-reflected light from the sample under investigation. Since it was first established in 1991 by Huang et al. [24], the clinical applications of OCT have been greatly improved. Ophthalmology remains the dominant one, initially applied in 1993 [17, 41]. The main reason is that OCT has limited penetration depth in biological tissues, but high resolution. The theory of OCT has been analyzed in details in review papers [14, 16, 32, 36, 44] in book chapters [15, 19, 42] and in books [4, 5, 10].

To derive a mathematical model for the OCT system, the scattering properties of the sample need to be described. There exist several different approaches to model the propagation of light within the sample: the radiative transfer equation with scattering and absorption coefficients [9, 38, 45], Lambert–Beer's law with the attenuation coefficient [39, 46], the equations of geometric optics with the refractive index [7], and Maxwell's equations with the susceptibility (or the refractive index) as optical parameters of the medium [6, 12, 27, 37, 43]. Also statistical approaches using Monte Carlo simulations are used [2, 11, 26, 31, 40].

This chapter describes the propagation of the electromagnetic wave through the sample using Maxwell's equations and adopts the analysis based on the theory of electromagnetic fields scattered by inhomogeneous media [8, 20]. The sample is hereby considered as a linear dielectric medium (potentially inhomogeneous and anisotropic). Moreover, the medium is considered weakly scattering so that the first-order Born approximation can be used and, as it is usually assumed in OCT, the backscattered light is detected far enough from the sample so that the far field approximation is valid. Starting from this model, different reconstruction formulas for special cases regarding the inner structure of the sample are presented.

This chapter is organized as follows. In Sect. 2, the principles of OCT and different variants of OCT systems are presented. Section 3 describes the solution of Maxwell's equations and an appropriate formula for the measurements of OCT is derived. Given the initial field and the optical properties (the susceptibility) of the

sample, the solution of the direct problem is obtained in Sect. 4. An iterative scheme is derived in the last section for the reconstruction of the unknown susceptibility, which is the inverse problem of OCT.

2 Basic Principles of OCT

OCT is used to gain information about the light scattering properties of an object by illuminating it with some short laser pulse and measuring the backscattered light.

The name "Optical Coherence Tomography" is motivated by the way the scattering data are measured: To get more precise measurements, the backscattered light is not directly detected, but first superimposed with the original laser pulse and then the intensity of this interference pattern is measured (this means that one measures the "coherence" of these two light beams).

Experimentally, this is done by separating the incoming light at a beam splitter into two identical beams which travel two different paths. One beam is simply reflected by a mirror and sent back to the beam splitter, while the other beam is directed to the sample. At the beam splitter, the beam reflected by the mirror and the backscattered light from the sample are recombined and sent to the detector [16, 25, 44]. See Fig. 1 for an illustration of this procedure.

There exist different variants of the OCT regarding the way the measurements are done:

Time and frequency domain OCT: In time domain OCT, the position of the mirror is varied and for each position one measurement is performed. On the other hand, in frequency domain OCT, the reference mirror is fixed and the detector is replaced by a spectrometer. Both methods provide equivalent measurements which are connected by a Fourier transform.

Standard and full field OCT: In standard OCT, the incoming light is focused through objective lenses to one spot in a certain depth in the sample and the backscattered light is measured in a point detector. This means that to obtain information of the whole sample, a transversal-lateral scan has to be performed

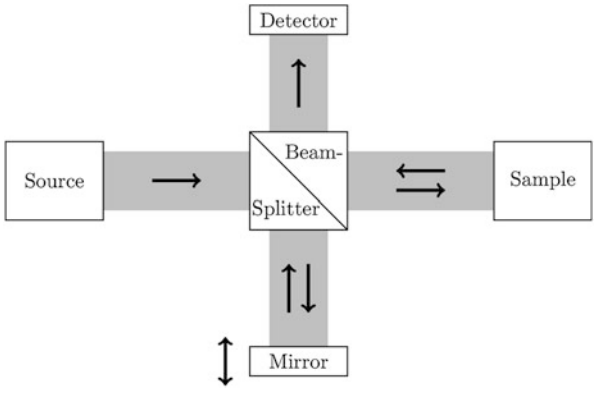


Fig. 1 Schematic diagram of the light traveling in an OCT system. The laser beam emitted by the source is divided at the beam splitter into two light beams; one is reflected at a mirror, the other one backscattered from the sample. The superposition of the two reflected beams is then measured at the detector

(by moving the light beam over the frontal surface of the sample). In full field OCT, the entire frontal surface of the sample is illuminated at once and the single point detector is replaced by a two-dimensional detector array, for instance by a charge-coupled device (CCD) camera.

Polarization-sensitive OCT: In classical OCT setups, the electromagnetic wave is simply treated as a scalar quantity. In polarization-sensitive OCT, however, the illuminating light beams are polarized and the detectors measure the intensity of the two polarization components of the interfered light.

There are also further modifications such as Doppler OCT and quantum OCT, which are not addressed here. In this chapter, the focus is mainly on time domain full field OCT, but also the others are discussed.

3 The Direct Scattering Problem

To derive a mathematical model for an OCT system, one has to describe on one hand the propagation and the scattering of the laser beam in the presence of the sample and on the other hand the way how this scattered wave is measured at the detectors. For the first part, the interaction of the incoming light with the sample can be modeled with Maxwell's macroscopic equations.

Maxwell's Equations

Maxwell's equations in matter consist of the partial differential equations

$$\operatorname{div}_x D(t,x) = 4\pi\rho(t,x), \qquad t \in \mathbb{R},\ x \in \mathbb{R}^3, \tag{1a}$$

$$\operatorname{div}_x B(t,x) = 0, \qquad t \in \mathbb{R},\ x \in \mathbb{R}^3, \tag{1b}$$

$$\operatorname{curl}_x E(t,x) = -\frac{1}{c}\frac{\partial B}{\partial t}(t,x), \qquad t \in \mathbb{R},\ x \in \mathbb{R}^3, \tag{1c}$$

$$\operatorname{curl}_x H(t,x) = \frac{4\pi}{c} J(t,x) + \frac{1}{c}\frac{\partial D}{\partial t}(t,x), \qquad t \in \mathbb{R},\ x \in \mathbb{R}^3, \tag{1d}$$

relating the following physical quantities (at some time $t \in \mathbb{R}$ and some location $x \in \mathbb{R}^3$):

Speed of light	c	$\mathbb{R}$
External charge density	$\rho(t,x)$	$\mathbb{R}$
External electric current density	$J(t,x)$	$\mathbb{R}^3$
Electric field	$E(t,x)$	$\mathbb{R}^3$
Electric displacement	$D(t,x)$	$\mathbb{R}^3$
Magnetic induction	$B(t,x)$	$\mathbb{R}^3$
Magnetic field	$H(t,x)$	$\mathbb{R}^3$

Maxwell's equations do not yet completely describe the propagation of the light (even assuming that the charge density ρ and the current density J are known, there are only 8 equations for the 12 unknowns E, D, B, and H).

Additionally to Maxwell's equations, it is therefore necessary to specify the relations between the fields D and E as well as between B and H.

Let $\Omega \subset \mathbb{R}^3$ denote the domain where the sample is located. It is considered as a nonmagnetic, dielectric medium without external charges or currents, this means that for all $t \in \mathbb{R}$ and all $x \in \Omega$ the electric and magnetic fields fulfil the relations

$$D(t,x) = E(t,x) + \int_0^\infty \chi(\tau,x)E(t-\tau,x)\mathrm{d}\tau, \tag{2a}$$

$$B(t,x) = H(t,x), \tag{2b}$$

$$\rho(t,x) = 0, \tag{2c}$$

$$J(t,x) = 0, \tag{2d}$$

where the function $\chi : \mathbb{R} \times \mathbb{R}^3 \to \mathbb{R}^{3\times 3}$ (for convenience, χ is also defined for negative times by $\chi(t,x) = 0$ for $t < 0$, $x \in \mathbb{R}^3$) is called the (electric) susceptibility and is the quantity to be recovered. The time dependence of χ hereby describes the fact that a change in the electric field E cannot immediately cause a change in the electric displacement D. Since this delay is quite small, it is sometimes ignored and $\chi(t,x)$ is then replaced by $\delta(t)\chi(x)$. Moreover, the medium is often considered to be isotropic, which means that χ is a multiple of the identity matrix.

The sample is situated in vacuum and the assumptions (2) are modified by setting for all $t \in \mathbb{R}$ and all $x \in \mathbb{R}^3 \setminus \Omega$

$$D(t,x) = E(t,x), \tag{3a}$$

$$B(t,x) = H(t,x), \tag{3b}$$

$$\rho(t,x) = 0, \tag{3c}$$

$$J(t,x) = 0. \tag{3d}$$

This simply corresponds to extend the Eq. (2) to $\mathbb{R} \times \mathbb{R}^3$ and to assume $\chi(t,x) = 0$ for all $t \in \mathbb{R}$, $x \in \mathbb{R}^3 \setminus \Omega$.

In this case of a nonmagnetic medium, Maxwell's equations result into one equation for the electric field E. To get rid of the convolution in (2a), it is practical to consider the Fourier transform with respect to time. In the following, the convention

$$\hat{f}(\omega,x) = \int_{-\infty}^{\infty} f(t,x)\mathrm{e}^{\mathrm{i}\omega t}\mathrm{d}t,$$

for the Fourier transform of a function f with respect to t is used.

Proposition 1. *Let E, D, B, and H fulfil Maxwell's equations (1). Moreover, let assumptions (2) and (3) be satisfied. Then the Fourier transform $\hat{E}$ of E fulfils the vector Helmholtz equation*

$$\operatorname{curl}_x \operatorname{curl}_x \hat{E}(\omega, x) - \frac{\omega^2}{c^2}(\mathbb{1} + \hat{\chi}(\omega, x))\hat{E}(\omega, x) = 0, \quad \omega \in \mathbb{R},\ x \in \mathbb{R}^3. \tag{4}$$

Proof. Applying the curl to (1c) and using (1d) with the assumptions $B = H$ and $J = 0$, yields

$$\operatorname{curl}_x \operatorname{curl}_x E(t, x) = -\frac{1}{c}\frac{\partial \operatorname{curl}_x B}{\partial t}(t, x) = -\frac{1}{c^2}\frac{\partial^2 D}{\partial t^2}(t, x). \tag{5}$$

The Fourier transform of (2a) and (3a) and the Fourier convolution theorem (recall that χ is set to zero outside Ω) imply that

$$\hat{D}(\omega, x) = (\mathbb{1} + \hat{\chi}(\omega, x))\hat{E}(\omega, x), \quad \text{for all} \quad \omega \in \mathbb{R},\ x \in \mathbb{R}^3.$$

Therefore, the Eq. (4) follows by taking the Fourier transform of (5). □

Initial Conditions

The sample is illuminated with a laser beam described initially (before it interacts with the sample) by the electric field $E^{(0)} : \mathbb{R} \times \mathbb{R}^3 \to \mathbb{R}^3$ which is (together with some magnetic field) a solution of Maxwell's equations (1) with the assumptions (3) for all $x \in \mathbb{R}^3$. Then, it follows from the proof of the Proposition 1, for $\chi = 0$, that

$$\operatorname{curl}_x \operatorname{curl}_x \hat{E}^{(0)}(\omega, x) - \frac{\omega^2}{c^2}\hat{E}^{(0)}(\omega, x) = 0, \quad \omega \in \mathbb{R},\ x \in \mathbb{R}^3. \tag{6}$$

Moreover, it is assumed that $E^{(0)}$ does not interact with the sample until the time $t = 0$, which means that $\operatorname{supp} E^{(0)}(t, \cdot) \cap \Omega = \emptyset$ for all $t \leq 0$.

The electric field $E : \mathbb{R} \times \mathbb{R}^3 \to \mathbb{R}^3$ generated by this incoming light beam in the presence of the sample is then a solution of Maxwell's equations (1) with the assumptions (2) and the initial condition

$$E(t, x) = E^{(0)}(t, x) \quad \text{for all} \quad t \leq 0,\ x \in \mathbb{R}^3. \tag{7}$$

Since Maxwell's equations for E in Proposition 1 are reformulated as an equation for the Fourier transform $\hat{E}$, it is helpful to rewrite the initial condition in terms of $\hat{E}$.

Proposition 2. *Let E (together with some magnetic field H) fulfil Maxwell's equations (1) with the assumptions (2) and (3) and with the initial condition (7).*

Then the Fourier transform of $E - E^{(0)}$ fulfils that the function $\omega \mapsto \hat{E}(\omega, x) - \hat{E}^{(0)}(\omega, x)$, defined on $\mathbb{R}$, can be extended to a square integrable, holomorphic function on the upper half plane $\mathbb{H} = \{\omega \in \mathbb{C} \mid \Im\mathrm{m}(\omega) > 0\}$ for every $x \in \mathbb{R}^3$.

Proof. From the initial condition (7) it follows that $E(t, x) - E^{(0)}(t, x) = 0$ for all $t \leq 0$. Thus, the result is a direct consequence form the Paley–Wiener theorem, which is based on the fact that in this case

$$\hat{E}(\omega, x) - \hat{E}^{(0)}(\omega, x) = \int_0^\infty (E - E^{(0)})(t, x) \mathrm{e}^{\mathrm{i}\omega t} \mathrm{d}t$$

is well defined for all $\omega \in \mathbb{H}$ and complex differentiable with respect to $\omega \in \mathbb{H}$. □

Remark that the electric field E is uniquely defined by (4) and Proposition 2.

The Measurements

The measurements are obtained by the combination of the backscattered field from the sample and the back-reflected field from the mirror. In practice, see Fig. 1, the sample and the mirror are in different positions. However, without loss of generality, a placement of them around the origin is assumed in the proposed formulation, in order to avoid rotating the coordinate system. To do so, the simultaneously illumination of the sample and the mirror is suppressed and two different illumination schemes are considered. The gain is to keep the same coordinate system but the reader should not be confused with illumination at different times.

Thus, the electric field E, which is obtained by illuminating the sample with the initial field $E^{(0)}$ (that is E solves (4) with the initial condition (7)), is combined with E_r which is the electric field obtained by replacing the sample by a mirror and illuminating with the same initial field $E^{(0)}$.

The mirror is placed orthogonal to the unit vector $e_3 = (0, 0, 1)$ through the point re_3. As in (7), it is assumed that supp $E^{(0)}(t, \cdot)$ does not interact with the mirror for $t < 0$, so that

$$E_r(t, x) = E^{(0)}(t, x) \quad \text{for all} \quad t < 0,\ x \in \mathbb{R}^3. \tag{8}$$

Then the resulting electric field $E_r : \mathbb{R} \times \mathbb{R}^3 \to \mathbb{R}^3$ is given as the solution of the same equations as E (Maxwell's equations (1) together with the assumptions (2) and initial condition (8)) with the susceptibility χ replaced by the susceptibility χ_r of the mirror at position r. One sort of (ideal) mirror can be described via the susceptibility $\chi_r(t, x) = 0$ for $x_3 > r$ and $\chi_r(t, x) = C\delta(t)\mathbb{1}$ for $x_3 \leq r$ with an (infinitely) large constant $C > 0$.

The intensity I_r of each component of the superposition of the electric fields E and E_r averaged over all time is measured at some detector points. The detectors are positioned at all points on the plane

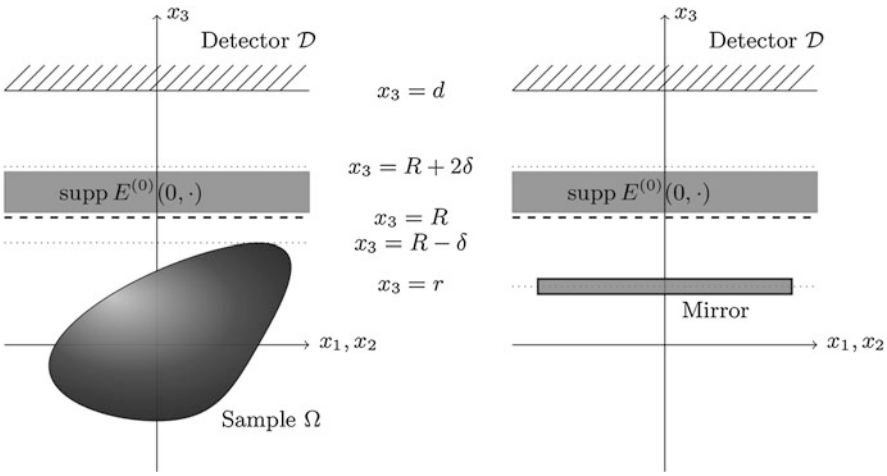


Fig. 2 The two scattering problems involved in OCT: On the left hand side the scattering of the initial wave on the sample Ω; on the right hand side the reference problem where the initial wave $E^{(0)}$ is reflected by a perfect mirror at a tunable position $r \in (-\infty, R)$. The two resulting electric fields, E and E_r, are then combined and this superposition $E + E_r$ is measured at the detector surface $\mathcal{D}$

$$\mathcal{D} = \{x \in \mathbb{R}^3 \mid x_3 = d\}$$

parallel to the mirror at a distance $d > 0$ from the origin. The mirror and the sample are both located in the lower half plane of the detector surface with some minimal distance to $\mathcal{D}$. Moreover, the highest possible position $R \in (\delta, d - 2\delta)$ of the mirror shall be by some distance $\delta > 0$ closer to the detector than the sample, this means (see Fig. 2)

$$\sup_{x \in \Omega} x_3 < R - \delta \quad \text{and} \quad r \in (-\infty, R). \tag{9}$$

To simplify the argument, let us additionally assume that the incoming electric field $E^{(0)}$ does not influence the detector after the time $t = 0$, meaning that

$$E^{(0)}(t, x) = 0 \quad \text{for all} \quad t \geq 0,\ x \in \mathcal{D}. \tag{10}$$

At the detector array, the data are obtained by measuring

$$I_{r,j}(x) = \int_0^\infty |E_j(t, x) + E_{r,j}(t, x)|^2 \mathrm{d}t, \quad x \in \mathcal{D},\ j \in \{1, 2, 3\}. \tag{11}$$

In standard OCT, the polarization is usually ignored. In this case, only the total intensity $I_r = \sum_{j=1}^3 I_{r,j}$ needs to be measured, see Sect. 5 for the reconstruction formulas in the isotropic case.

In this measurement setup, it is easy to acquire besides the intensity I_r also the intensity of the two waves E and E_r separately by blocking one of the two waves E and E_r at a time. Practically, it is sometimes not even necessary to measure them since the intensity of the reflected laser beam E_r can be explicitly calculated from the knowledge of the initial beam $E^{(0)}$, and the intensity of E is usually negligible compared with the intensity I_r (because of the assumption (10), the field E contains only backscattered light at the detector after the measurement starts). Therefore, one can consider instead of I_r the function

$$M_{r,j}(x) = \frac{1}{2}\left(I_{r,j} - \int_0^\infty |E_j(t,x)|^2 \mathrm{d}t - \int_0^\infty |E_{r,j}(t,x)|^2 \mathrm{d}t\right) \tag{12}$$

for $r \in (-\infty, R)$, $j \in \{1,2,3\}$, and $x \in \mathcal{D}$ as the measurement data.

Proposition 3. *Let the initial conditions (7) and (8) and the additional assumption (10) be satisfied. Then, for all $x \in \mathcal{D}$, $r \in (-\infty, R)$, and $j \in \{1,2,3\}$ the measurements M_r, defined by (12), fulfil*

$$M_{r,j}(x) = \int_{-\infty}^{\infty} (E_j - E_j^{(0)})(t,x)(E_{r,j} - E_j^{(0)})(t,x)\mathrm{d}t \tag{13a}$$

$$= \int_{-\infty}^{\infty} (\hat{E}_j - \hat{E}_j^{(0)})(\omega,x)\overline{(\hat{E}_{r,j} - \hat{E}_j^{(0)})(\omega,x)}\mathrm{d}\omega, \tag{13b}$$

Proof. Expanding the function $I_{r,j}$, given by (11), gives

$$I_{r,j}(x) = \int_0^\infty \left(|E_j(t,x)|^2 + |E_{r,j}(t,x)|^2 + 2E_j(t,x)E_{r,j}(t,x)\right)\mathrm{d}t.$$

Thus, by the Definition (12) of M_r, it follows that

$$M_{r,j}(x) = \int_0^\infty E_j(t,x)E_{r,j}(t,x)\mathrm{d}t,$$

which, using the assumption (10), can be rewritten in the form

$$M_{r,j}(x) = \int_0^\infty (E_j - E_j^{(0)})(t,x)(E_{r,j} - E_j^{(0)})(t,x)\mathrm{d}t.$$

Then, since E and E_r coincide with $E^{(0)}$ for $t < 0$, see (7) and (8), the integration can be extended to all times. This proves the formula (13a) for M_r. The second formula follows from Plancherel's theorem. □

4 Solution of the Direct Problem

In this section the solution of the direct problem, to determine the measurements M_r, defined by (13a), from the susceptibility χ, is derived using Born and far field approximation for the electric field.

Proposition 4. *Let E be a solution of the Eqs. (4) and (7). Then, the Fourier transform $\hat{E}$ solves the Lippmann–Schwinger integral equation*

$$\hat{E}(\omega,x) = \hat{E}^{(0)}(\omega,x) + \left(\frac{\omega^2}{c^2}\mathbb{1} + \mathrm{grad}_x \mathrm{div}_x\right) \int_{\mathbb{R}^3} G(\omega, x-y)\hat{\chi}(\omega,y)\hat{E}(\omega,y)\mathrm{d}y, \tag{14}$$

where G is the fundamental solution of the Helmholtz equation given by

$$G(\omega,x) = \frac{\mathrm{e}^{\mathrm{i}\frac{\omega}{c}|x|}}{4\pi|x|}, \quad x \neq 0,\ \omega \in \mathbb{R}.$$

Proof. Equation (4) can be rewritten in the form

$$\mathrm{curl}_x \mathrm{curl}_x \hat{E}(\omega,x) - \frac{\omega^2}{c^2}\hat{E}(\omega,x) = \phi(\omega,x)$$

with the inhomogeneity

$$\phi(\omega,x) = \frac{\omega^2}{c^2}\hat{\chi}(\omega,x)\hat{E}(\omega,x). \tag{15}$$

Using that $\hat{E}^{(0)}$ solves (6), the difference $\hat{E} - \hat{E}^{(0)}$ satisfies the inhomogeneous vector Helmholtz equation

$$\mathrm{curl}_x \mathrm{curl}_x (\hat{E} - \hat{E}^{(0)})(\omega,x) - \frac{\omega^2}{c^2}(\hat{E} - \hat{E}^{(0)})(\omega,x) = \phi(\omega,x). \tag{16}$$

The divergence of this equation, using that $\mathrm{div}_x \mathrm{curl}_x (\hat{E} - \hat{E}^{(0)}) = 0$, implies

$$\mathrm{div}_x (\hat{E} - \hat{E}^{(0)})(\omega,x) = -\frac{c^2}{\omega^2}\mathrm{div}_x \phi(\omega,x). \tag{17}$$

Applying the vector identity

$$\mathrm{curl}_x \mathrm{curl}_x (\hat{E} - \hat{E}^{(0)}) = \mathrm{grad}_x \mathrm{div}_x (\hat{E} - \hat{E}^{(0)}) - \Delta_x (\hat{E} - \hat{E}^{(0)})$$

in (16) and using (17) yields

$$\Delta_x(\hat{E}-\hat{E}^{(0)})(\omega,x)+\frac{\omega^2}{c^2}(\hat{E}-\hat{E}^{(0)})(\omega,x)=-\frac{c^2}{\omega^2}\mathrm{grad}_x\mathrm{div}_x\phi(\omega,x)-\phi(\omega,x).$$

This is a Helmholtz equation for $\hat{E}-\hat{E}^{(0)}$ and the general solution which is (with respect to ω) holomorphic in the upper half plane (equivalent to (7) by Proposition 2) is given by, see [8]

$$\begin{aligned}(\hat{E}-\hat{E}^{(0)})(\omega,x)&=-\frac{c^2}{\omega^2}\int_{\mathbb{R}^3}G(\omega,x-y)\left(\frac{\omega^2}{c^2}\mathbb{1}+\mathrm{grad}_y\mathrm{div}_y\right)\phi(\omega,y)\mathrm{d}y\\&=-\frac{c^2}{\omega^2}\left(\frac{\omega^2}{c^2}\mathbb{1}+\mathrm{grad}_x\mathrm{div}_x\right)\int_{\mathbb{R}^3}G(\omega,x-y)\phi(\omega,y)\mathrm{d}y.\end{aligned}$$

For the last equality, integration by parts and $\mathrm{grad}_x G(\omega,x-y)=-\mathrm{grad}_y G(\omega,x-y)$ were used. The Lippmann–Schwinger equation (14) follows from the last expression inserting the expression (15) for ϕ. □

This integral equation uniquely defines the electric field E. The reader is referred to [1,8] for the isotropic case and to [33] for an anisotropic medium.

Born and Far Field Approximation

To solve the Lippmann–Schwinger equation (14), the medium is assumed to be weakly scattering, which means that $\hat{\chi}$ is sufficiently small (implying that the difference $E-\hat{E}^{(0)}$ becomes small compared to $E^{(0)}$) so that the Born approximation $E^{(1)}$, defined by

$$\begin{aligned}\hat{E}^{(1)}(\omega,x)=\hat{E}^{(0)}(\omega,x)+\left(\frac{\omega^2}{c^2}\mathbb{1}+\mathrm{grad}_x\mathrm{div}_x\right)\\\int_{\mathbb{R}^3}G(\omega,x-y)\hat{\chi}(\omega,y)\hat{E}^{(0)}(\omega,y)\mathrm{d}y,\end{aligned}\tag{18}$$

is considered a good approximation for the electric field E, see [3]. To describe multiple scattering events, one considers higher order Born approximations. For different linearization techniques, the reader is referred to [1,23]. Moreover, since the detector in OCT is typically quite far away from the sample, one can simplify the expression (18) for the electric field at the detector array by replacing it with its asymptotic behavior for $|x|\to\infty$, that is replace the formula for $E^{(1)}$ by its far field approximation (the far field approximation could also be applied to the solution E of the Lippmann–Schwinger equation (14)).

Proposition 5. *Consider, for a given function $\phi:\mathbb{R}^3\to\mathbb{R}^3$ with compact support and some parameter $a\in\mathbb{R}$, the function*

$$g : \mathbb{R}^3 \to \mathbb{R}^3, \quad g(x) = \int_{\mathbb{R}^3} \frac{\mathrm{e}^{\mathrm{i}a|x-y|}}{|x-y|} \phi(y) \mathrm{d}y.$$

Then, it follows, asymptotically for $\rho \to \infty$ and uniformly in $\vartheta \in S^2$, that

$$(a^2 + \mathrm{grad}_x \mathrm{div}_x) g(\rho\vartheta) \simeq -\frac{a^2 \mathrm{e}^{\mathrm{i}a\rho}}{\rho} \int_{\mathbb{R}^3} \vartheta \times (\vartheta \times \phi(y)) \mathrm{e}^{-\mathrm{i}a\langle \vartheta, y \rangle} \mathrm{d}y \tag{19}$$

Proof. Consider the function

$$\Gamma : \mathbb{R}^3 \to \mathbb{C}, \quad \Gamma(x) = \frac{\mathrm{e}^{\mathrm{i}a|x|}}{|x|}.$$

Then

$$\begin{aligned} \frac{\partial^2 \Gamma}{\partial x_j \partial x_k}(x) &= \frac{\partial}{\partial x_j} \left[\left(\frac{\mathrm{i}a}{|x|^2} - \frac{1}{|x|^3} \right) x_k \mathrm{e}^{\mathrm{i}a|x|} \right] \\ &= \left[\left(\frac{\mathrm{i}a}{|x|^2} - \frac{1}{|x|^3} \right) \delta_{jk} + \left(\frac{\mathrm{i}a}{|x|^2} - \frac{1}{|x|^3} \right) \frac{\mathrm{i}a x_j x_k}{|x|} \right. \\ &\quad \left. + \left(-2 \frac{\mathrm{i}a}{|x|^3} + 3 \frac{1}{|x|^4} \right) \frac{x_j x_k}{|x|} \right] \mathrm{e}^{\mathrm{i}a|x|}. \end{aligned}$$

Therefore, writing x in spherical coordinates: $x = \rho\vartheta$ with $\rho > 0$, $\vartheta \in S^2$, for $\rho \to \infty$ uniformly in ϑ, it can be seen that

$$\frac{\partial^2 \Gamma}{\partial x_j \partial x_k}(\rho\vartheta) = -\frac{a^2 \mathrm{e}^{\mathrm{i}a\rho}}{\rho} \vartheta_j \vartheta_k + \mathcal{O}\left(\frac{1}{\rho^2} \right),$$

The approximation (locally uniformly in $y \in \mathbb{R}^3$)

$$|\rho\vartheta - y| = \rho \sqrt{|\vartheta|^2 - \frac{2}{\rho} \langle \vartheta, y \rangle + \frac{1}{\rho^2} |y|^2} = \rho - \langle \vartheta, y \rangle + \mathcal{O}\left(\frac{1}{\rho} \right),$$

implies that (again uniformly in $\vartheta \in S^2$)

$$\frac{\partial^2 \Gamma}{\partial x_j \partial x_k}(\rho\vartheta - y) = -\frac{a^2 \mathrm{e}^{\mathrm{i}a(\rho - \langle \vartheta, y \rangle)}}{\rho} \vartheta_j \vartheta_k + \mathcal{O}\left(\frac{1}{\rho} \right).$$

Now, considering the compact support of ϕ and using that $x \in \mathbb{R}^3 \setminus \mathrm{supp}\phi$

$$(\mathrm{grad}_x \mathrm{div}_x g)_j(x) = \sum_{k=1}^{3} \frac{\partial}{\partial x_j} \int_{\mathbb{R}^3} \frac{\partial \Gamma}{\partial x_k}(x - y) \phi_k(y) \mathrm{d}y$$

$$= \int_{\mathbb{R}^3} \sum_{k=1}^{3} \frac{\partial^2 \Gamma}{\partial x_j \, \partial x_k}(x-y)\phi_k(y)\mathrm{d}y.$$

Asymptotically for $|x| \to \infty$ (again using the compact support of ϕ) one obtains

$$a^2 g_j(\rho\vartheta) + (\mathrm{grad}_x \mathrm{div}_x g)_j(\rho\vartheta) \simeq a^2 \int_{\mathbb{R}^3} \sum_{k=1}^{3} \frac{\mathrm{e}^{\mathrm{i}a(\rho - \langle \vartheta, y\rangle)}}{\rho} \left(\delta_{jk} - \vartheta_j \vartheta_k\right) \phi_k(y)\mathrm{d}y.$$

The approximation (19) follows from the vector identity $\vartheta \times (\vartheta \times \phi) = \langle \vartheta, \phi \rangle \, \vartheta - |\vartheta|^2 \phi$ and $|\vartheta| = 1$. □

The application of both the far field and the Born approximation, this means Proposition 5 for the expression (18) of $E^{(1)}$, that is setting $a = \omega/c$ and $\phi = \frac{1}{4\pi} \hat{\chi} \hat{E}^{(0)}$ in Proposition 5, imply the asymptotic behavior

$$\hat{E}^{(1)}(\omega, \rho\vartheta) \simeq \hat{E}^{(0)}(\omega, \rho\vartheta) - \frac{\omega^2 \mathrm{e}^{\mathrm{i}\frac{\omega}{c}\rho}}{4\pi\rho c^2} \int_{\mathbb{R}^3} \vartheta \times \left(\vartheta \times (\hat{\chi}(\omega, y)\hat{E}^{(0)}(\omega, y))\right) \mathrm{e}^{-\mathrm{i}\frac{\omega}{c}\langle \vartheta, y\rangle} \mathrm{d}y. \tag{20}$$

The Forward Operator

To obtain a forward model for the measurements described in Sect. 3, the (approximative) formula (20) is considered as a model for the solution of the scattering problem. To make this formula concrete, one has to plug in a function $E^{(0)}$ describing the initial illumination (recall that $E^{(0)}$ has to solve (6)).

The specific illumination is a laser pulse propagating in the direction $-e_3$, orthogonal to the detector surface $\mathcal{D} = \{x \in \mathbb{R}^3 \mid x_3 = d\}$, this means

$$E^{(0)}(t, x) = f(t + \tfrac{x_3}{c})p, \tag{21}$$

which solves Maxwell's equations (1) with the assumptions (3) for some fixed vector $p \in \mathbb{R}^3$, with $p_3 = \langle p, e_3 \rangle = 0$, describing the polarization of the initial laser beam.

Proposition 6. *The function $E^{(0)}$, defined by (21) with $\langle p, e_3 \rangle = 0$, solves together with the magnetic field $H^{(0)}$, defined by*

$$H^{(0)}(t, x) = f(t + \tfrac{x_3}{c}) p \times e_3,$$

Maxwell's equations (1) in the vacuum, that is with the additional assumptions (3).

Proof. The four equations of (1) can be directly verified:

$$\operatorname{div}_x E^{(0)}(t,x) = \frac{1}{c} f'(t + \tfrac{x_3}{c}) \langle e_3, p \rangle = 0,$$
$$\operatorname{div}_x H^{(0)}(t,x) = \frac{1}{c} f'(t + \tfrac{x_3}{c}) \langle e_3, p \times e_3 \rangle = 0,$$
$$\operatorname{curl}_x E^{(0)}(t,x) = \frac{1}{c} f'(t + \tfrac{x_3}{c}) e_3 \times p = -\frac{1}{c} \frac{\partial H^{(0)}}{\partial t}(t,x),$$
$$\operatorname{curl}_x H^{(0)}(t,x) = \frac{1}{c} f'(t + \tfrac{x_3}{c}) e_3 \times (p \times e_3) = \frac{1}{c} f'(t + \tfrac{x_3}{c}) p = \frac{1}{c} \frac{\partial E^{(0)}}{\partial t}(t,x).$$

□

To guarantee that the initial field $E^{(0)}$ (and also the magnetic field $H^{(0)}$) does not interact with the sample or the mirror for $t \leq 0$ and neither contributes to the measurement at the detectors for $t \geq 0$ as required by (8) and (10) the vertical distribution $f : \mathbb{R} \to \mathbb{R}$ should satisfy (see Fig. 2)

$$\operatorname{supp} f \subset (\tfrac{R}{c}, \tfrac{d}{c}). \tag{22}$$

In the case of an illumination $E^{(0)}$ of the form (21), the electric field E_r produced by an ideal mirror at the position r is given by

$$E_r(t,x) = \begin{cases} \big(f(t + \frac{x_3}{c}) - f(t + \frac{x_3}{c} + 2\,\frac{r - x_3}{c})\big) p & \text{if } x_3 > r, \\ 0 & \text{if } x_3 \leq r. \end{cases} \tag{23}$$

This just corresponds to the superposition of the initial wave with the (orthogonally) reflected wave, which travels additionally the distance $2\frac{x_3 - r}{c}$. The change in polarization of the reflected wave (from p to $-p$) comes from the fact that the tangential components of the electric field have to be continuous across the border of the mirror.

The following proposition gives the form of the measurements M_r, described in Sect. 3, on the detector surface $\mathcal{D}$ for the specific illumination (21).

Proposition 7. *Let $E^{(0)}$ be an initial illumination of the form (21) satisfying (22). Then, the equations for the measurements M_r from Proposition 3 are given by*

$$M_{r,j}(x) = -p_j \int_{-\infty}^{\infty} (E_j - E_j^{(0)})(t,x) f(t + \tfrac{2r - x_3}{c}) \mathrm{d}t, \tag{24a}$$
$$= -\frac{p_j}{2\pi} \int_{-\infty}^{\infty} (\hat{E}_j - \hat{E}_j^{(0)})(\omega, x) \hat{f}(-\omega) \mathrm{e}^{\mathrm{i}\frac{\omega}{c}(2r - x_3)} \mathrm{d}\omega \tag{24b}$$

for all $j \in \{1,2,3\}$, $r \in (-\infty, R)$, and $x \in \mathcal{D}$.

Proof. Since the electric field E_r reflected on a mirror at vertical position $r \in (-\infty, R)$ is according to (23) given by

$$E_r(t,x) = \big(f(t + \tfrac{x_3}{c}) - f(t + \tfrac{2r-x_3}{c})\big)p \quad \text{for all} \quad t \in \mathbb{R},\ x \in \mathcal{D},$$

the measurement functions M_r (defined by (12) and computed with (13a)) are simplified, for the particular initial illumination $E^{(0)}$ of the form (21), to (24a) for $x \in \mathcal{D}$.

Since formula (24a) is just a convolution, the electric field $E - E^{(0)}$ can be rewritten, in terms of its Fourier transform, in the form

$$M_{r,j}(x) = -\frac{p_j}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (\hat{E}_j - \hat{E}_j^{(0)})(\omega, x) \mathrm{e}^{-\mathrm{i}\omega t} f(t + \tfrac{2r-x_3}{c}) \mathrm{d}\omega \mathrm{d}t.$$

Interchanging the order of integration and applying the Fourier transform $\hat{f}$ of f, it follows Eq. (24b). □

In the limiting case of a delta impulse as initial wave, that is for $f(\xi) = \delta(\xi - \xi_0)$ with some constant $\xi_0 \in (\frac{R}{c}, \frac{d}{c})$ satisfying (22), the measurements provide directly the electric field. Indeed, it can be seen from (24a) that

$$M_{r,j}(x) = -p_j (E_j - E_j^{(0)})(\tfrac{x_3-2r}{c} + \xi_0, x).$$

By varying $r \in (-\infty, R)$, the electric field E can be obtained (to be more precise, its component in direction of the initial polarization) as a function of time at every detector position.

The following assumptions are made:

Assumption 5. *The susceptibility χ is sufficiently small so that the Born approximation $E^{(1)}$ for the solution E of the Lippmann–Schwinger equation (14) can be applied.*

Assumption 6. *The detectors are sufficiently far away from the object so that one can use the far field asymptotics (20) for the measured field.*

Under these assumptions, one can approximate the electric field by the far field expression of the Born approximation $E^{(1)}$ and plug in the expression in (20) to obtain the measurements $M_{r,j}$, $j \in \{1,2,3\}$.

The above analysis, introducing appropriate operators, can then be formulated as an operator equation. The integral equation (20) can be formally written as

$$(\hat{E}^{(1)} - \hat{E}^{(0)})(\omega, x) = (\mathcal{K}_0 \hat{\chi})(\omega, x),$$

for a given $\hat{E}^{(0)}$ where the operator $\mathcal{K}_0 : \hat{\chi} \mapsto \hat{E}^{(1)} - \hat{E}^{(0)}$ is given by

$$(\mathcal{K}_0 v)(\omega, \rho\vartheta) = -\frac{\omega^2 \mathrm{e}^{\mathrm{i}\frac{\omega}{c}\rho}}{4\pi\rho c^2} \int_{\mathbb{R}^3} \vartheta \times \big(\vartheta \times (v(\omega, y)\hat{E}^{(0)}(\omega, y))\big)\mathrm{e}^{-\mathrm{i}\frac{\omega}{c}\langle\vartheta, y\rangle}\mathrm{d}y,$$
$$\rho > 0, \vartheta \in S^2.$$

It is emphasized that $\hat{\chi} : \mathbb{R} \times \mathbb{R}^3 \to \mathbb{C}^{3\times 3}$ and $\hat{E}^{(1)} - \hat{E}^{(0)} : \mathbb{R} \times \mathbb{R}^3 \to \mathbb{C}^3$, that is $\mathcal{K}_0 v$ is a function from $\mathbb{R} \times \mathbb{R}^3$ into $\mathbb{C}^3$. Equivalently, considering the Eq. (13b), one has

$$M(r, x) = \big(M_{r,j}(x)\big)_{j=1}^3 = \big(\mathcal{M}(\hat{E}^{(1)} - \hat{E}^{(0)})\big)(r, x),$$

where the operator $\mathcal{M}$ is defined by

$$(\mathcal{M}v)(r, x) = \left(\int_{-\infty}^{\infty} v_j(\omega, x)\overline{(\hat{E}_{r,j} - \hat{E}_j^{(0)})(\omega, x)}\mathrm{d}\omega\right)_{j=1}^3, \quad x \in \mathcal{D}.$$

Here, $\mathcal{M}v$ is a function from $\mathbb{R} \times \mathcal{D}$ to $\mathbb{R}^3$. Thus, combining the operators $\mathcal{K}_0$ and $\mathcal{M}$, the forward operator $\mathcal{F} : \hat{\chi} \mapsto M$, $\mathcal{F} = \mathcal{M}\mathcal{K}_0$ models the direct problem. The inverse problem of OCT is then formulated as an operator equation

$$\mathcal{F}\hat{\chi} = M. \tag{25}$$

For the specific illumination (21), one has

$$\hat{E}^{(0)}(\omega, x) = \left(\int_{-\infty}^{\infty} f(t + \tfrac{x_3}{c})\mathrm{e}^{\mathrm{i}\omega t}\mathrm{d}t\right) p = \hat{f}(\omega)\mathrm{e}^{-\mathrm{i}\frac{\omega}{c}x_3} p. \tag{26}$$

Then, the operators $\mathcal{K}_0$ and $\mathcal{M}$ simplify to

$$(\mathcal{K}_0 v)(\omega, \rho\vartheta) = -\frac{\omega^2 \mathrm{e}^{\mathrm{i}\frac{\omega}{c}\rho}}{4\pi\rho c^2}\hat{f}(\omega) \int_{\mathbb{R}^3} \vartheta \times \big(\vartheta \times (v(\omega, y)\, p)\big)\mathrm{e}^{-\mathrm{i}\frac{\omega}{c}\langle\vartheta + e_3, y\rangle}\mathrm{d}y \tag{27}$$

and, recalling that $M_{r,3} = 0$ since $p_3 = 0$ (the polarization in the incident direction is zero),

$$(\mathcal{M}v)(r, x) = \left(-\frac{p_j}{2\pi}\int_{-\infty}^{\infty} v_j(\omega, x)\,\hat{f}(-\omega)\mathrm{e}^{\mathrm{i}\frac{\omega}{c}(2r - x_3)}\mathrm{d}\omega\right)_{j=1}^2. \tag{28}$$

The operator $\mathcal{K}_0$ is derived from the Born approximation taking into account the far field approximation for the solution of the Lippmann–Schwinger equation (14). But, one could also neglect Assumption 5 and Assumption 6 and use the operator $\mathcal{K}$ corresponding to Eq. (14), that is,

$$(\mathcal{K}v)(\omega, x) = \left(\frac{\omega^2}{c^2}\mathbb{1} + \mathrm{grad}_x \mathrm{div}_x\right) \int_{\mathbb{R}^3} G(\omega, x-y) v(\omega, y) \hat{E}(\omega, y) \mathrm{d}y,$$

and considering the nonlinear forward operator $\mathcal{F} = \mathcal{M}\mathcal{K}$.

The next section focuses on the solution of (25), considering the operators $\mathcal{K}_0$ and $\mathcal{M}$, given by (27) and (28), respectively. The inversion of $\mathcal{F}$ is performed in two steps, first $\mathcal{M}$ is inverted and then $\mathcal{K}_0$.

5 The Inverse Scattering Problem

In optical coherence tomography, the susceptibility χ of the sample is imaged from the measurements $M_r(x)$, $r \in (-\infty, R)$, $x \in \mathcal{D}$. In a first step, it is shown that the measurements allow us to reconstruct the scattered field on the detector $\mathcal{D}$, that is inverting the operator (28).

The vertical distribution $f : \mathbb{R} \to \mathbb{R}$ should additionally satisfy (see Fig. 2)

$$\mathrm{supp} f \subset (\tfrac{R}{c}, \tfrac{R}{c} + \tfrac{2\delta}{c}) \subset (\tfrac{R}{c}, \tfrac{d}{c}) \quad \text{for some} \quad \delta > 0. \tag{29}$$

This guarantees that the initial field $E^{(0)}$ (and also the magnetic field $H^{(0)}$) not interact with the sample or the mirror for $t \leq 0$ and neither contribute to the measurement at the detectors for $t \geq 0$ as required by (8) and (10).

The condition that the length of the support of $E^{(0)}$ is at most 2δ (the assumption that the support starts at $\frac{R}{c}$ is only made to simplify the notation) is required for Proposition 8. It ensures that the formula (13a) for the measurement data $M_r(x)$, $x \in \mathcal{D}$, vanishes for values $r \geq R$ so that the integral on the right hand side of (30) is only over the interval $(-\infty, R)$ where measurement data are obtained (recall that measurements are only performed for positions $r < R$ of the mirror).

Proposition 8. *Let $E^{(0)}$ be an initial illumination of the form (21) satisfying (29). Then, the measurements M_r from Proposition 7 imply for the electric field E:*

$$(\hat{E}_j - \hat{E}_j^{(0)})(\omega, x)\overline{\hat{f}(\omega)} p_j = -\frac{2}{c} \int_{-\infty}^{R} M_{r,j}(x) \mathrm{e}^{-\mathrm{i}\frac{\omega}{c}(2r - x_3)} \mathrm{d}r \tag{30}$$

for all $j \in \{1, 2, 3\}$, $\omega \in \mathbb{R}$, and $x \in \mathcal{D}$.

Proof. Remark that the formula (24a) can be extended to all $r \in \mathbb{R}$ by setting $M_{r,j}(x) = 0$ for $r \geq R$. Indeed, from (29) it follows that $E(t, \cdot) = E^{(0)}(t, \cdot)$ for all $t < \frac{\delta}{c}$. Since E is a solution of the linear wave equation with constant wave speed c on the half space given by $x_3 > R - \delta$, the difference between E and $E^{(0)}$ caused by the sample needs at least time $\frac{d-R+\delta}{c}$ to travel from the point at $x_3 = R - \delta$ to the detector at $x_3 = d$, so:

$$E(t, x) = E^{(0)}(t, x) \quad \text{for all} \quad x \in \mathbb{R}^3 \text{ with } x_3 = d \text{ and } ct < 2\delta + d - R.$$

This means that the integrand vanishes for $t < \frac{2\delta+d-R}{c}$. In the case of $t \geq \frac{2\delta+d-R}{c}$, it holds for $r \geq R$ that

$$ct + 2r - d \geq 2\delta + d - R + 2R - d = R + 2\delta,$$

so that $f(t + \frac{2r-d}{c}) = 0$ by the assumption (29) on the support of f. Therefore, for $r \geq R$, always one of the factors in the integrand in (24a) is zero which implies that $M_r(x) = 0$ for $r \geq R$ and $x \in \mathcal{D}$.

Thus, Eq. (24b) holds for all $r \in \mathbb{R}$ and applying the inverse Fourier transform with respect to r, using that $\hat{f}(-\omega) = \overline{\hat{f}(\omega)}$ because f is real valued, yields

$$\frac{2}{c}\int_{-\infty}^{\infty} M_{r,j}(x)\mathrm{e}^{-\mathrm{i}\frac{2\omega r}{c}}\mathrm{d}r = -p_j(\hat{E}_j - \hat{E}_j^{(0)})(\omega, x)\overline{\hat{f}(\omega)}\mathrm{e}^{-\mathrm{i}\frac{\omega}{c}x_3},$$

which can equivalently be written as (30). □

This means that one can calculate from the Fourier transform of the measurements $r \mapsto M_r(x)$ at some frequency ω the Fourier transform of the electric field at ω as long as the Fourier transform of the initial wave $E^{(0)}$ does not vanish at ω, that is for $\hat{f}(\omega) \neq 0$. Thus, under the Assumption 5 and Assumption 6, Eq. (30) can be solved for the electric field $\hat{E}$. Proposition 8 thus provides the inverse of the operator $\mathcal{M}$ defined by (28). Now, the inversion of the operator $\mathcal{K}_0$ given by (27) is performed considering the optical properties of the sample.

Proposition 9. *Let $E^{(0)}(t, x)$ be given by the form (21) with $p_3 = 0$ and the additional assumption (29). Then, for every $\omega \in \mathbb{R} \setminus \{0\}$ with $\hat{f}(\omega) \neq 0$, the formula*

$$p_j\big[\vartheta \times (\vartheta \times \tilde{\chi}(\omega, \tfrac{\omega}{c}(\vartheta + e_3))p)\big]_j \simeq \frac{8\pi\rho c}{\omega^2|\hat{f}(\omega)|^2}\int_{-\infty}^{R} M_{r,j}(\rho\vartheta)\mathrm{e}^{-\mathrm{i}\frac{\omega}{c}(2r-\rho(\vartheta_3-1))}\mathrm{d}r \tag{31}$$

holds for all $j \in \{1, 2\}$, $\vartheta \in S^2_+ := \{\eta \in S^2 \mid \eta_3 > 0\}$, and $\rho = \frac{d}{\vartheta_3}$ (asymptotically for $\chi \to 0$ and $\rho \to \infty$).

Here $\tilde{\chi}$ denotes the Fourier transform of χ with respect to time and space, that is

$$\tilde{\chi}(\omega, k) = \int_{-\infty}^{\infty}\int_{\mathbb{R}^3} \chi(t, x)\mathrm{e}^{-\mathrm{i}\langle k,x\rangle}\mathrm{e}^{\mathrm{i}\omega t}\mathrm{d}x\mathrm{d}t = \int_{\mathbb{R}^3} \hat{\chi}(\omega, x)\mathrm{e}^{-\mathrm{i}\langle k,x\rangle}\mathrm{d}x. \tag{32}$$

Proof. Because of (26), the Fourier transform of the electric field $E - E^{(0)}$ can be approximated, using (20) with $E \simeq E^{(1)}$ (by Assumption 5 and Assumption 6), by

$$(\hat{E} - \hat{E}^{(0)})(\omega, \rho\vartheta) \simeq -\frac{\omega^2\hat{f}(\omega)\mathrm{e}^{\mathrm{i}\frac{\omega}{c}\rho}}{4\pi\rho c^2}\int_{\mathbb{R}^3} \mathrm{e}^{-\mathrm{i}\frac{\omega}{c}(\langle y,\vartheta\rangle+y_3)}\vartheta \times (\vartheta \times \hat{\chi}(\omega, y)p)\mathrm{d}y.$$

Then, applying (32), one obtains

$$(\hat{E} - \hat{E}^{(0)})(\omega, \rho\vartheta) \simeq -\frac{\omega^2 \hat{f}(\omega) \mathrm{e}^{\mathrm{i}\frac{\omega}{c}\rho}}{4\pi\rho c^2} \vartheta \times (\vartheta \times \tilde{\chi}(\omega, \tfrac{\omega}{c}(\vartheta + e_3))p). \tag{33}$$

From (30), it is known that, for $\hat{f}(\omega) \neq 0$ and $p_j \neq 0$,

$$(\hat{E}_j - \hat{E}_j^{(0)})(\omega, \rho\vartheta) = -\frac{2}{p_j c \overline{\hat{f}(\omega)}} \int_{-\infty}^{R} M_{r,j}(\rho\vartheta) \mathrm{e}^{-\mathrm{i}\frac{\omega}{c}(2r - \rho\vartheta_3)} \mathrm{d}r.$$

This identity together with (33), asymptotically for $\omega \neq 0$, yields the statement (31). □

To derive reconstruction formulas, Proposition 9 is used, which states that from the measurements M_r (under the Assumption 5 and Assumption 6) the expression

$$p_j \big[\vartheta \times (\vartheta \times \tilde{\chi}(\omega, \tfrac{\omega}{c}(\vartheta + e_3))p)\big]_j, \quad j = 1, 2, \tag{34}$$

can be calculated. Here, $p \in \mathbb{R}^2 \times \{0\}$ denotes the polarization of the initial illumination $E^{(0)}$, see (21), and $\vartheta \in S_+^2$ is the direction from the origin (where the sample is located) to a detector.

The Isotropic Case

This section analyzes the special case of an isotropic medium, meaning that the susceptibility matrix χ is just a multiple of the unit matrix, so in the following χ is identified with a scalar.

Then, from the sum of the measurements $M_{r,1}$ and $M_{r,2}$, using the formula (31), one obtains the expression

$$\tilde{\chi}(\omega, \tfrac{\omega}{c}(\vartheta + e_3)) \langle p, \vartheta \times (\vartheta \times p)\rangle = \tilde{\chi}(\omega, \tfrac{\omega}{c}(\vartheta + e_3))(\langle \vartheta, p\rangle^2 - |p|^2).$$

Since $\langle \vartheta, p\rangle^2 < |p|^2$ for every combination of $p \in \mathbb{R}^2 \times \{0\}$ and $\vartheta \in S_+^2$, one has direct access to the spatial and temporal Fourier transform

$$\tilde{\chi}(\omega, \tfrac{\omega}{c}(\vartheta + e_3)), \quad \omega \in \mathbb{R} \setminus \{0\}, \ \vartheta \in S_+^2, \tag{35}$$

of χ in a subset of $\mathbb{R} \times \mathbb{R}^3$.

However, it remains the problem of reconstructing the four-dimensional susceptibility data χ from the three-dimensional measurement data (35). In the following, some different additional assumptions are discussed to compensate the lack of dimension, see Table 1.

Table 1 Different assumptions about the susceptibility and the corresponding reconstruction formulas

Assumptions	Reconstruction method	Section
$\tilde{\chi}(\omega, k) = \tilde{\chi}(k)$	Reconstruction from partial (three dimensional) Fourier data: $\tilde{\chi}(k), k \in \mathbb{R}^3$, $\measuredangle(k, e_3) \in (-\frac{\pi}{4}, \frac{\pi}{4})$	Non-dispersive Medium in Full Field OCT
$\tilde{\chi}(\omega, k) = \tilde{\chi}(k_3)$	Reconstruction from full (one dimensional) Fourier data: $\tilde{\chi}(k_3), k_3 \in \mathbb{R} \setminus \{0\}$	Non-dispersive Medium with Focused Illumination
supp $\chi(\cdot, x) \subset [0, T]$ $\mathcal{R}(\chi(\tau, \cdot))(\cdot, \varphi)$ is piecewise constant	Recursive formula to get limited angle Radon data $\mathcal{R}(\chi(\tau, \cdot))(\sigma, \varphi), \sigma \in \mathbb{R}, \varphi \in S^2$ with $\measuredangle(\varphi, e_3) \in (-\frac{\pi}{4}, \frac{\pi}{4})$	Dispersive Medium
$\chi(\tau, x) = \delta(x_1)\delta(x_2)\chi(\tau, x_3)$, supp $\chi(\cdot, x) \subset [0, T]$, and $\chi(\tau, \cdot)$ is piecewise constant	Recursive formula to reconstruct χ	Dispersive Layered Medium with Focused Illumination

Here, $(\mathcal{R}g)(\sigma, \varphi) = \int_{\{x \in \mathbb{R}^3 \mid \langle x, \varphi \rangle = \sigma\}} g(y) \mathrm{d}s(y)$, $\sigma \in \mathbb{R}, \varphi \in S^2$, denotes the Radon transform of a function $g : \mathbb{R}^3 \to \mathbb{R}$.

Non-dispersive Medium in Full Field OCT

The model is simplified by assuming an immediate reaction of the sample to the exterior electric field in (2a). This means that χ can be considered as a delta distribution in time so that its temporal Fourier transform $\hat{\chi}$ does not depend on frequency, that is $\hat{\chi}(\omega, x) = \hat{\chi}(x)$. Thus, the reconstruction reduces to the problem of finding $\hat{\chi}$ from its partial (spatial) Fourier data

$$\begin{aligned}\tilde{\chi}(k) \quad \text{for} \quad k \in \{\tfrac{\omega}{c}(\vartheta + e_3) \in \mathbb{R}^3 \mid \vartheta \in S^2_+,\ \omega \in \mathbb{R} \setminus \{0\}\} \\ = \{\kappa \in \mathbb{R}^3 \setminus \{0\} \mid \arccos(\langle \tfrac{\kappa}{|\kappa|}, e_3 \rangle) \in (-\tfrac{\pi}{4}, \tfrac{\pi}{4})\}.\end{aligned}$$

Thus, only the Fourier data of χ in the right circular cone $\mathcal{C}$ with axis along e_3 and aperture $\frac{\pi}{2}$ are observed (see Fig. 3). In practice, these data are usually only available for a small range of frequencies ω.

Inverse scattering for full field OCT, under the Born approximation, has been considered by Marks et al. [28, 29] where algorithms to recover the scalar susceptibility were proposed.

Non-dispersive Medium with Focused Illumination

In standard OCT, the illumination is focused to a small region inside the object so that the function χ can be assumed to be constant in the directions e_1 and e_2 (locally the illumination is still assumed to be properly described by a plane wave). Then, the problem can be reduced by two dimensions assuming that the illumination is described by a delta distribution in these two directions. As before, χ is assumed to be frequency independent, so that

$$\hat{\chi}(\omega, x) = \delta(x_1)\delta(x_2)\hat{\chi}(x_3),$$

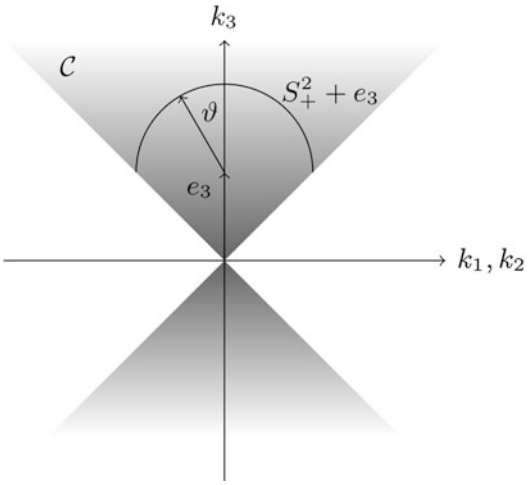


Fig. 3 Region $\mathcal{C} \cong \mathbb{R} \times (S^2_+ + e_3)$ of the available Fourier data of χ

this means that the spatial and temporal Fourier transform (32) fulfils $\tilde{\chi}(\omega, k) = \tilde{\chi}(k_3)$. In this case, the two-dimensional detector array can be replaced by a single point detector located at de_3.

Then, the measurement data (35) in direction $\vartheta = e_3$ provide the Fourier transform

$$\tilde{\chi}(\tfrac{2\omega}{c}) \quad \text{for all} \quad \omega \in \mathbb{R} \setminus \{0\}.$$

Therefore, the reconstruction of the (one dimensional) susceptibility $x_3 \mapsto \hat{\chi}(x_3)$ can be simply obtained by an inverse Fourier transform.

This one-dimensional analysis, has been used initially by Fercher et al. [18], reviewed in [13] and by Hellmuth [22] to describe time domain OCT. Ralston et al. [34,35] described the OCT system using a single backscattering model. The solution was given through numerical simulation using regularized least squares methods.

Dispersive Medium

However, in the case of a dispersive medium, that is frequency-dependent $\hat{\chi}$, the difficulty is to reconstruct the four-dimensional function $\chi : \mathbb{R} \times \mathbb{R}^3 \to \mathbb{C}$ from the three-dimensional data

$$\hat{m} : \mathbb{R} \times S^2_+ \to \mathbb{C}, \quad \hat{m}(\omega, \vartheta) = \tilde{\chi}(\omega, \tfrac{\omega}{c}(\vartheta + e_3)). \tag{36}$$

Lemma 1. *Let $\hat{m}$ be given by (36). Then its inverse Fourier transform $m : \mathbb{R} \times S^2_+ \to \mathbb{C}$ with respect to the first variable is given by*

$$m(t, \vartheta) = \frac{c}{\sqrt{2(1+\vartheta_3)}} \int_{-\infty}^{\infty} \bar{\chi}(\tau; \tau - t, \vartheta) \mathrm{d}\tau, \quad t \in \mathbb{R}, \ \vartheta \in S^2_+, \tag{37}$$

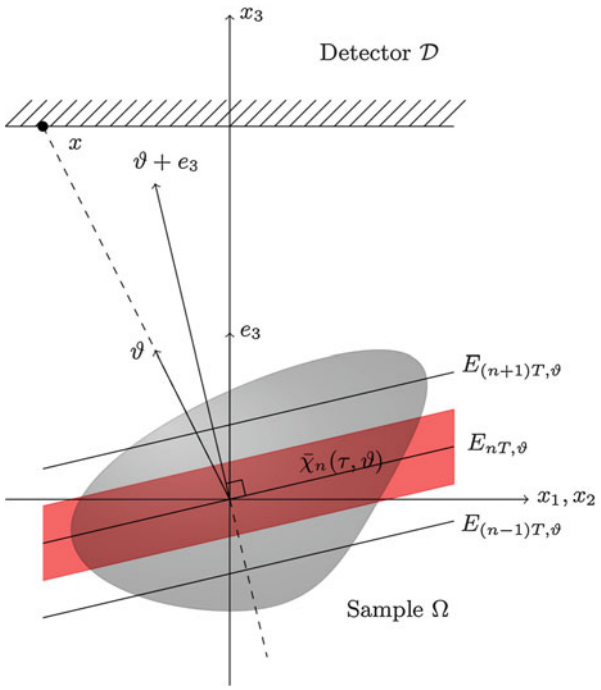


Fig. 4 Discretization of χ with respect to the detection points

where

$$\bar{\chi}(\tau;\sigma,\vartheta) = \int_{E_{\sigma,\vartheta}} \chi(\tau, y)\mathrm{d}s(y), \quad \tau,\sigma \in \mathbb{R},\ \vartheta \in S^2_+,$$

and $E_{\sigma,\vartheta}$ *denotes the plane*

$$E_{\sigma,\vartheta} = \{y \in \mathbb{R}^3 \mid \langle \vartheta + e_3, y\rangle = c\sigma\}, \quad \sigma \in \mathbb{R},\ \vartheta \in S^2_+. \tag{38}$$

Proof. Taking the inverse temporal Fourier transform of $\hat{m}$ and using (32), it follows that

$$\begin{aligned} m(t,\vartheta) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \tilde{\chi}(\omega, \tfrac{\omega}{c}(\vartheta + e_3))\mathrm{e}^{-\mathrm{i}\omega t}\mathrm{d}\omega \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{\mathbb{R}^3} \hat{\chi}(\omega, x)\mathrm{e}^{-\mathrm{i}\frac{\omega}{c}\langle\vartheta+e_3,x\rangle}\mathrm{e}^{-\mathrm{i}\omega t}\mathrm{d}x\mathrm{d}\omega. \end{aligned}$$

Interchanging the order of integration, the integral over ω is again described by an inverse Fourier transform and the previous equation becomes

$$m(t,\vartheta) = \int_{\mathbb{R}^3} \chi(t + \tfrac{1}{c}\langle \vartheta + e_3, x\rangle, x)\mathrm{d}x.$$

Substituting then the variable x_3 by $\tau = t + \frac{1}{c}\langle \vartheta + e_3, x\rangle$, this can be written as

$$m(t,\vartheta) = \frac{c}{1+\vartheta_3}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \chi(\tau, \psi_{\tau-t,\vartheta}(x_1,x_2))\mathrm{d}x_1\mathrm{d}x_2\mathrm{d}\tau \tag{39}$$

with the function

$$\psi_{\sigma,\vartheta} : \mathbb{R}^2 \to \mathbb{R}^3, \quad \psi_{\sigma,\vartheta}(x_1,x_2) = \left(x_1, x_2, \frac{c\sigma}{1+\vartheta_3} - \frac{\vartheta_1 x_1 + \vartheta_2 x_2}{1+\vartheta_3}\right).$$

Now, $\psi_{\sigma,\vartheta}$ is seen to be the parametrization of the plane

$$E_{\sigma,\vartheta} = \{y \in \mathbb{R}^3 \mid \langle v_\vartheta, y\rangle = a_{\sigma,\vartheta}\} \quad \text{with} \quad v_\vartheta = \begin{pmatrix} \frac{\vartheta_1}{1+\vartheta_3} \\ \frac{\vartheta_2}{1+\vartheta_3} \\ 1 \end{pmatrix} \quad \text{and} \quad a_{\sigma,\vartheta} = \frac{c\sigma}{1+\vartheta_3},$$

see Fig. 4. The square root of the Gram determinant of the parametrization $\psi_{\sigma,\vartheta}$ is now given by the length of the vector $v_\vartheta = \frac{\partial\psi_{\sigma,\vartheta}}{\partial x_1} \times \frac{\partial\psi_{\sigma,\vartheta}}{\partial x_2}$, which implies that

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \chi(\tau, \psi_{\tau-t,\vartheta}(x_1,x_2))\mathrm{d}x_1\mathrm{d}x_2 = \sqrt{\frac{1+\vartheta_3}{2}}\int_{E_{\tau-t,\vartheta}} \chi(\tau,y)\mathrm{d}s(y).$$

Plugging this into (39) yields the claim. □

Thus, the measurements give the combination (37) of values $\bar{\chi}$ of the Radon transform of the function $\chi(\tau,\cdot)$. It seems, however, impossible to recover the values $\bar{\chi}(\tau;\sigma,\vartheta)$ from this combination

$$m(t,\vartheta) = \frac{c}{\sqrt{2(1+\vartheta_3)}}\int_{-\infty}^{\infty} \bar{\chi}(\tau;\tau-t,\vartheta)\mathrm{d}\tau,$$

since (for every fixed angle $\vartheta \in S^2_+$) one would have to reconstruct a function on $\mathbb{R}^2$ from one dimensional data.

To overcome this problem, the function $\bar{\chi}(\tau;\cdot,\vartheta)$ is going to be discretized for every $\tau \in \mathbb{R}$ and $\vartheta \in S^2_+$, where the step size will depend on the size of the support of $\chi(\cdot,x)$.

Let us therefore consider the following assumption.

Assumption 7. *The support of χ in the time variable is contained in a small interval $[0,T]$ for some $T > 0$:*

$$\operatorname{supp}\chi(\cdot, x) \subset [0, T] \quad \textit{for all} \quad x \in \mathbb{R}^3.$$

Then, the following discretization

$$\bar{\chi}_n(\tau, \vartheta) = \int_{E_{nT,\vartheta}} \chi(\tau, y) \mathrm{d}s(y), \quad n \in \mathbb{Z},\ \tau \in (0, T),\ \vartheta \in S^2_+,$$

of the Radon transform of the functions $\chi(\tau, \cdot)$ is considered, where $E_{\sigma,\vartheta}$ denotes the plane defined in (38).

Assumption 8. *The value $\bar{\chi}_n(\tau, \vartheta)$ is a good approximation for the integral of the function $\chi(\tau, \cdot)$ over the planes $E_{nT+\varepsilon,\vartheta}$ for all $\varepsilon \in [-\frac{T}{2}, \frac{T}{2})$ (see Fig. 4), that is*

$$\bar{\chi}_n(\tau, \vartheta) \approx \int_{E_{nT+\varepsilon,\vartheta}} \chi(\tau, y) \mathrm{d}s(y), \quad \varepsilon \in [-\tfrac{T}{2}, \tfrac{T}{2}),\ n \in \mathbb{Z},\ \tau \in (0, T),\ \vartheta \in S^2_+.$$

Under the Assumption 8, Eq. (37) can be rewritten in the form

$$m(t, \vartheta) \approx \frac{c}{\sqrt{2(1 + \vartheta_3)}} \int_0^T \bar{\chi}_{N(\tau - t)}(\tau, \vartheta) \mathrm{d}\tau,$$

where $N(\sigma) = \lfloor \frac{\sigma}{T} + \frac{1}{2} \rfloor$ denotes the integer closest to $\frac{\sigma}{T}$. This (approximate) identity can now be iteratively solved for $\bar{\chi}$.

Proposition 10. *Let*

$$\bar{m}(t, \vartheta) = \frac{c}{\sqrt{2(1 + \vartheta_3)}} \int_0^T \bar{\chi}_{N(\tau - t)}(\tau, \vartheta) \mathrm{d}\tau, \quad \vartheta \in S^2_+,\ t \in \mathbb{R},$$

for some constant $T > 0$ with the integer-valued function $N(\sigma) = \lfloor \frac{\sigma}{T} + \frac{1}{2} \rfloor$.

Then, $\bar{\chi}$ fulfils the recursion relation

$$\bar{\chi}_n(\tau, \vartheta) = \bar{\chi}_{n+1}(\tau, \vartheta) + \frac{\sqrt{2(1 + \vartheta_3)}}{c} \frac{\partial \bar{m}}{\partial t}(\tau - (n + \tfrac{1}{2})T, \vartheta),$$
$$n \in \mathbb{Z},\ \tau \in (0, T),\ \vartheta \in S^2_+. \tag{40}$$

Proof. Let $t = -nT + \varepsilon$ with $\varepsilon \in [-\frac{T}{2}, \frac{T}{2})$ and $n \in \mathbb{Z}$, then

$$N(\tau - t) = \begin{cases} n & \text{if } \tau \in (0, \frac{T}{2} + \varepsilon), \\ n + 1 & \text{if } \tau \in [\frac{T}{2} + \varepsilon, T). \end{cases}$$

Taking an arbitrary $\varepsilon \in [-\frac{T}{2}, \frac{T}{2})$ and $n \in \mathbb{Z}$, $\bar{m}$ can be formulated as

$$\bar{m}(-nT+\varepsilon,\vartheta)=\frac{c}{\sqrt{2(1+\vartheta_3)}}\left(\int_0^{\frac{T}{2}+\varepsilon}\bar{\chi}_n(\tau,\vartheta)\mathrm{d}\tau+\int_{\frac{T}{2}+\varepsilon}^{T}\bar{\chi}_{n+1}(\tau,\vartheta)\mathrm{d}\tau\right).$$

Differentiating this equation with respect to ε, it follows that

$$\frac{\partial\bar{m}}{\partial t}(-nT+\varepsilon,\vartheta)=\frac{c}{\sqrt{2(1+\vartheta_3)}}\left(\bar{\chi}_n(\tfrac{T}{2}+\varepsilon,\vartheta)-\bar{\chi}_{n+1}(\tfrac{T}{2}+\varepsilon,\vartheta)\right),$$

which (with $\tau=\frac{T}{2}+\varepsilon$) is equivalent to (40). □

Thus, given that $\bar{\chi}_n(t,\vartheta)=0$ for sufficiently large $n\in\mathbb{Z}$ (recall that supp $\chi(\tau,\cdot)\subset\Omega$ for all $\tau\in(0,T)$), one can recursively reconstruct $\bar{\chi}$, to obtain the data

$$\int_{E_{\sigma,\vartheta}}\chi(\tau,y)\mathrm{d}s(y)\quad\text{for all}\quad\tau\in[0,T),\ \sigma\in\mathbb{R},\ \vartheta\in S^2_+$$

for the Radon transform of $\chi(\tau,\cdot)$.

However, since the plane $E_{\sigma,\vartheta}$ is by its Definition (38) orthogonal to the vector $\vartheta+e_3$ for $\vartheta\in S^2_+$, this provides only the values of the Radon transform corresponding to planes which are orthogonal to a vector in the cone $\mathcal{C}$, see Fig. 3. For the reconstruction, one therefore still has to invert a limited angle Radon transform.

Dispersive Layered Medium with Focused Illumination

Except from ophthalmology, OCT is also widely used for investigation of skin deceases, such as cancer. From the mathematical point of view, this simplifies the main model since the human skin can be described as a multilayer structure with different optical properties and varying thicknesses in each layer.

Here the incident field is considered to propagate with normal incidence to the interface $x_3=L$ and the detector array is replaced by a single point detector located at de_3. The susceptibility is simplified as

$$\chi(t,x)=\delta(x_1)\delta(x_2)\chi(t,x_3),$$

and therefore the measurements provide the data, see (37) with $\tilde{\chi}(\omega,k)=\tilde{\chi}(\omega,k_3)$,

$$\hat{m}(\omega)=\tilde{\chi}(\omega,\tfrac{\omega}{c}2e_3),\quad\omega\in\mathbb{R}\setminus\{0\}.$$

Considering the special structure of a layered medium, the susceptibility is described by a piecewise constant function in x_3. This means explicitly that χ has the form

$$\chi(t,x_3)=\begin{cases}\chi_0:=0, & x_3\notin[0,L]\\ \chi_n(t), & x_3\in[L_n,L_{n+1})\end{cases},\quad n=1,\dots,N \tag{41}$$

with (unknown) parameters $L = L_1 > L_2 > \ldots > L_{N+1} = 0$ characterizing the thicknesses of the N layers and (unknown) functions χ_n.

Lemma 1, for $\vartheta = e_3$, gives

$$m(t) = \frac{c}{2}\int_{-\infty}^{\infty} \bar{\chi}(\tau; \tau - t)\mathrm{d}\tau, \quad \text{where} \quad \bar{\chi}(\tau;\sigma) = \chi(\tau, \tfrac{c\sigma}{2}).$$

Remarking that $\bar{\chi}$ is piecewise constant (41) and additionally assuming that $\chi(\cdot, x_3)$ has compact support, see Assumption 7, with $T < \frac{2}{c}\min_n(L_n - L_{n+1})$ Proposition 10 can be applied for $\vartheta = e_3$ to iteratively reconstruct χ starting from $\chi_0 = 0$.

Modified Born Approximation

In the proposed iteration scheme, Proposition 10, the traveling of the incident field through the sample before reaching a "specific" layer, where the susceptibility is to be reconstructed, is not considered. To do so, a modified iteration method is presented describing the traveling of the light through the different layers using Frensel's equations.

The main idea is to consider, for example, in the second step of the recursive formula, given χ_1 to find χ_2, as incident the field $\hat{E}^{(0)}$, given by (26), traveled also through the first layer. This process can be continued to the next steps.

Let us first introduce some notations which will be used in the following. The fields $\hat{E}_n^{(r)}$ and $\hat{E}_n^{(t)}$ denote the reflected and the transmitted fields, with respect to the boundary L_n, respectively. The transmitted field $\hat{E}_n^{(t)}$ after traveling through the n-th layer is incident on the L_{n+1} boundary and is denoted by $\hat{E}_{n+1}^{(0)}$. The reflected field by the L_{n+1} boundary back to the L_n boundary will be denoted by $\hat{E}_{n+1}^{(r)}$ and by $\hat{E}_n^{(r')}$ after traveling through the n-th layer (see Fig. 5). To simplify this model, multiple reflections are not included and the electric fields are taken to be tangential to the interface planes.

Lemma 2. *Let the sample have susceptibility given by (41) and let ρ_n and τ_n denote the reflection and the corresponding transmission coefficients for the L_n boundary, respectively. Then, the field incident on the n-th layer with respect to the initial incident field $\hat{E}^{(0)} := \hat{E}_1^{(0)}$ is given by*

$$\begin{pmatrix} \hat{E}_n^{(0)} \\ 0 \end{pmatrix} = (\mathcal{M}_1 \cdot \mathcal{M}_2 \cdot \ldots \cdot \mathcal{M}_{n-1})^{-1} \begin{pmatrix} \hat{E}^{(0)} \\ \hat{E}_1^{(r)} \end{pmatrix} \quad \textit{for} \quad n = N-1, \ldots, 2$$

assuming no backward field in the n-th layer, where

$$\mathcal{M}_n = \frac{1}{\tau_n}\begin{pmatrix} \mathrm{e}^{\mathrm{i}\frac{\omega}{c}\sqrt{\chi_n+1}(L_n - L_{n+1})} & \rho_n \mathrm{e}^{-\mathrm{i}\frac{\omega}{c}\sqrt{\chi_n+1}(L_n - L_{n+1})} \\ \rho_n \mathrm{e}^{\mathrm{i}\frac{\omega}{c}\sqrt{\chi_n+1}(L_n - L_{n+1})} & \mathrm{e}^{-\mathrm{i}\frac{\omega}{c}\sqrt{\chi_n+1}(L_n - L_{n+1})} \end{pmatrix}.$$

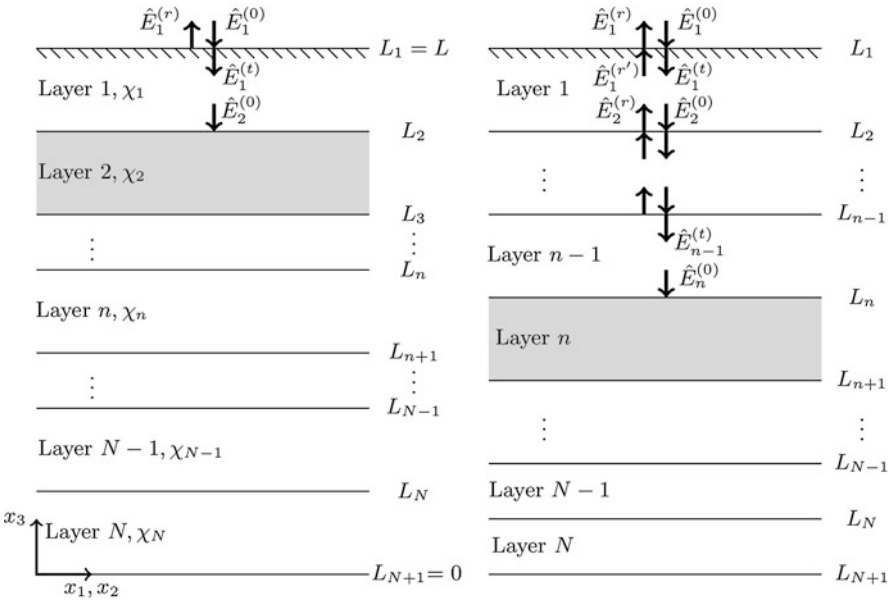


Fig. 5 Layered medium. Propagation of the initial field through the sample incident on the second layer (*left image*) and in general on the n-th layer, for $n = 3, \ldots, N$ (*right image*)

Proof. Because of the assumptions (normal incidence, $\hat{E}^{(0)}$ tangential to the boundary) the boundary conditions require the continuity of the total (upward and downward) electric and magnetic fields. Then, the reflection ρ_n and the corresponding transmission τ_n coefficients for the L_n boundary in terms of the susceptibility are given by [21]

$$\rho_n = \frac{\sqrt{\chi_{n-1}+1} - \sqrt{\chi_n + 1}}{\sqrt{\chi_{n-1}+1} + \sqrt{\chi_n + 1}}, \quad \tau_n = 1 + \rho_n.$$

To determine the propagation equations for the electric fields, the transfer matrices formulation is applied [30]. In particular, the fields at the top of the n-th layer can be computed with respect to the fields at the top of the $(n+1)$th using

$$\begin{pmatrix} \hat{E}_n^{(0)} \\ \hat{E}_n^{(r)} \end{pmatrix} = \mathcal{M}_n \begin{pmatrix} \hat{E}_{n+1}^{(0)} \\ \hat{E}_{n+1}^{(r)} \end{pmatrix} \quad \text{for} \quad n = N-1, \ldots, 1.$$

and with respect to the incident field,

$$\begin{pmatrix} \hat{E}_1^{(0)} \\ \hat{E}_1^{(r)} \end{pmatrix} = \mathcal{M}_1 \cdot \ldots \cdot \mathcal{M}_{n-1} \begin{pmatrix} \hat{E}_n^{(0)} \\ \hat{E}_n^{(r)} \end{pmatrix} \quad \text{for} \quad n = N-1, \ldots, 2.$$

□

From the previous result, given χ_n (by the recursion relation of Proposition 10), the matrix $\mathcal{M}_{n+1}$ is computed to obtain the update $\hat{E}^{(0)}_{n+1}$ which is then incident to the rest part of the sample. This means that $\hat{E}^{(0)}$ is replaced by $\hat{E}^{(0)}_{n+1}$ in the derivation of the measurements and the recursion relation (Lemma 1 and Proposition 10) for computing χ_{n+1}. For example, in the second step to reconstruct χ_2, the incident field is simply given by

$$\hat{E}^{(0)}_2 = \tau_1 \mathrm{e}^{-\mathrm{i}\frac{\omega}{c}\sqrt{\chi_1+1}(L_1-L_2)}\hat{E}^{(0)}.$$

The only unknown in this representation is the boundary L_2 which can be approximated considering the point where change in the value of the measured function $\bar{m}$ is observed. The following analysis can be also extended for anisotropic media, but in a more complicated context since the displacement D and the electric field E are not always parallel.

A simplification usually made here is to consider the sample field as the sum of all the discrete reflections and neglect dispersion. This mathematical model was adopted by Bruno and Chaubell [7] for solving the inverse scattering problem of determining the refractive index and the width of each layer from the output data. The solution was obtained using the Gauss–Newton method and the effect of the initial guesses was also considered.

In conclusion, the traveling of the scattered field from the n-th layer through the sample could also be considered. Since the spherical waves can be represented as a superposition of plane waves by using similar techniques, in a more complicated form, one can obtain the transmitted scattered field.

The Anisotropic Case

In the anisotropic case, the susceptibility χ cannot be considered a multiple of the identity. Therefore, the problem is to reconstruct from the expressions

$$p_j\big[\vartheta \times (\vartheta \times \tilde{\chi}(\omega, \tfrac{\omega}{c}(\vartheta + e_3))p)\big]_j, \quad j = 1, 2,$$

see (34), the matrix-valued function $\chi : \mathbb{R} \times \mathbb{R}^3 \to \mathbb{R}^{3\times 3}$, where it is assumed that measurements for every polarization $p \in \mathbb{R}^2 \times \{0\}$ of the initial field $E^{(0)}$ are available.

Introducing in analogy to (36) the function

$$\hat{m}_{p,j} : \mathbb{R} \times S^2_+ \to \mathbb{C}, \quad \hat{m}_{p,j}(\omega, \vartheta) = \tilde{\chi}_{\vartheta,p,j}(\omega, \tfrac{\omega}{c}(\vartheta + e_3)),$$

where $\tilde{\chi}_{\vartheta,p,j}$ is for every $\vartheta \in S^2_+$, $p \in \mathbb{R}^2 \times \{0\}$, and $j \in \{1, 2\}$ the (spatial and temporal) Fourier transform of

$$\chi_{\vartheta,p,j} : \mathbb{R} \times \mathbb{R}^3 \to \mathbb{R}, \quad \chi_{\vartheta,p,j}(t, x) = p_j\big[\vartheta \times (\vartheta \times \chi(t, x)p)\big]_j,$$

Lemma 1 (with m replaced by $m_{p,j}$ and χ replaced by $\chi_{\vartheta,p,j}$) can be applied to find that the inverse Fourier transform of $\hat{m}_{p,j}$ with respect to its first variable fulfils

$$m_{p,j}(t,\vartheta) = \frac{c}{\sqrt{2(1+\vartheta_3)}} \int_{-\infty}^{\infty} \int_{E_{\tau-t,\vartheta}} \chi_{\vartheta,p,j}(\tau, y) \mathrm{d}s(y) \mathrm{d}\tau.$$

Now, the same assumptions as in the isotropic case are considered, namely Assumption 7 and similar to Assumption 8:

Assumption 9. *The approximation*

$$\int_{E_{nT,\vartheta}} \chi_{\vartheta,p,j}(\tau, y) \mathrm{d}s(y) \approx \int_{E_{nT+\varepsilon,\vartheta}} \chi_{\vartheta,p,j}(\tau, y) \mathrm{d}s(y) \quad \textit{for all} \quad \varepsilon \in [-\tfrac{T}{2}, \tfrac{T}{2})$$

is for every $\tau \in \mathbb{R}$, $\vartheta \in S_+^2$, $n \in \mathbb{Z}$, $p \in \mathbb{R}^2 \times \{0\}$, *and* $j \in \{1,2\}$ *justified.*

Then, Proposition 10 provides an approximate reconstruction formula for the functions

$$\bar{\chi}_{p,j}(\tau;\sigma,\vartheta) = \int_{E_{\sigma,\vartheta}} \chi_{\vartheta,p,j}(\tau, y) \mathrm{d}s(y) = p_j \left[\vartheta \times (\vartheta \times \bar{\chi}(\tau;\sigma,\vartheta)\, p)\right]_j \tag{42}$$

for all $p \in \mathbb{R}^2 \times \{0\}$, $\tau \in \mathbb{R}$, $\sigma \in \mathbb{R}$, $\vartheta \in S_+^2$, and $j \in \{1,2\}$, where

$$\bar{\chi}(\tau;\sigma,\vartheta) = \int_{E_{\sigma,\vartheta}} \chi(\tau, y) \mathrm{d}s(y) \tag{43}$$

denotes the two-dimensional Radon transform data of the function $\chi(\tau, \cdot)$.

Proposition 11. *Let* $\vartheta \in S_+^2$ *be fixed and* $a_{p,j}$, $p \in \mathbb{R}^2 \times \{0\}$, $j = 1,2$, *be such that the equations*

$$p_j [\vartheta \times (\vartheta \times Xp)]_j = a_{p,j} \quad \textit{for all} \quad p \in \mathbb{R}^2 \times \{0\},\ j \in \{1,2\}, \tag{44}$$

for the matrix $X \in \mathbb{R}^{3\times 3}$ *have a solution.*

Then $X \in \mathbb{R}^{3\times 3}$ *is a solution of* (44) *if and only if*

$$(P_\vartheta X)_{k\ell} = B_{k\ell}, \quad B = \begin{pmatrix} -a_{e_1,1} & a_{e_1,1} - a_{e_1+e_2,1} \\ a_{e_2,2} - a_{e_1+e_2,2} & -a_{e_2,2} \end{pmatrix}, \quad k,\ell \in \{1,2\}, \tag{45}$$

where $P_\vartheta \in \mathbb{R}^{3\times 3}$ *denotes the orthogonal projection in direction* ϑ.

Proof. First, remark that the equation system (44) is equivalent to the four equations

$$\begin{aligned} a_{e_1,1} &= [\vartheta \times (\vartheta \times Xe_1)]_1, & a_{e_1+e_2,1} &= a_{e_1,1} + [\vartheta \times (\vartheta \times Xe_2)]_1, \\ a_{e_2,2} &= [\vartheta \times (\vartheta \times Xe_2)]_2, & a_{e_1+e_2,2} &= a_{e_2,2} + [\vartheta \times (\vartheta \times Xe_1)]_2, \end{aligned} \tag{46}$$

which correspond to the Eq. (44) for $(p, j) \in \{(e_1, 1), (e_2, 2), (e_1 + e_2, 1), (e_1 + e_2, 2)\}$. Indeed, for arbitrary polarization $p = p_1 e_1 + p_2 e_2$, the expression $p_j[\vartheta \times (\vartheta \times Xp)]_j$ can be written as a linear combination of the four expressions $[\vartheta \times (\vartheta \times Xe_i)]_k, i, k = 1, 2$:

$$\begin{aligned} p_1[\vartheta \times (\vartheta \times Xp)]_1 &= p_1^2[\vartheta \times (\vartheta \times Xe_1)]_1 + p_1 p_2[\vartheta \times (\vartheta \times Xe_2)]_1, \\ p_2[\vartheta \times (\vartheta \times Xp)]_2 &= p_1 p_2[\vartheta \times (\vartheta \times Xe_1)]_2 + p_2^2[\vartheta \times (\vartheta \times Xe_2)]_2, \end{aligned}$$

and is thus determined by (46).

Now, the equation system (46) written in matrix form reads

$$[\vartheta \times (\vartheta \times Xp)]_k = -\left[B \begin{pmatrix} p_1 \\ p_2 \end{pmatrix}\right]_k, \quad k \in \{1, 2\}, \tag{47}$$

for all $p \in \mathbb{R}^2 \times \{0\}$ with B defined by (45).

Decomposing $Xp = \langle \vartheta, Xp \rangle \vartheta + P_\vartheta Xp$, where $P_\vartheta \in \mathbb{R}^{3\times 3}$ denotes the orthogonal projection in direction ϑ, and using that

$$\vartheta \times (\vartheta \times Xp) = \vartheta \times (\vartheta \times P_\vartheta Xp) = \langle \vartheta, P_\vartheta Xp \rangle \vartheta - P_\vartheta Xp = -P_\vartheta Xp,$$

the Eq. (47) can be written in the form (45). □

Proposition 11 applied to the Eq. (42) for the matrix $X = \bar{\chi}(\tau; \sigma, \vartheta)$ for some fixed values $\tau, \sigma \in \mathbb{R}$ and $\vartheta \in S^2_+$ shows that the data $a_{p,j} = p_j[\vartheta \times (\vartheta \times \bar{\chi}(\tau; \sigma, \vartheta))]_j$ for $j = 1, 2$ and the three different polarization vectors $p = e_1$, $p = e_2$, and $p = e_1 + e_2$ uniquely determine with Eq. (45) the projection

$$(P_\vartheta \bar{\chi}(\tau; \sigma, \vartheta))_{k,\ell} = \int_{E_{\sigma,\vartheta}} (P_\vartheta \chi(\tau, y))_{k,\ell} \, \mathrm{d}s(y) \quad \text{for} \quad k, \ell \in \{1, 2\}.$$

Moreover, measurements for additional polarizations p do not provide any further information so that at every detector point, corresponding to a direction $\vartheta \in S^2_+$, only the four elements $(P_\vartheta \chi)_{k,\ell}$, $k, \ell = 1, 2$, of the projection $P_\vartheta \chi$ influence the measurements.

To obtain additional data which make a full reconstruction of χ possible, one can carry out extra measurements after slight rotations of the sample.

So, let $R \in \mathrm{SO}(3)$ describe the rotation of the sample. Then the transformed susceptibility χ_R is given by

$$\chi_R(t, y) = R\chi(t, R^{\mathrm{T}} y)R^{\mathrm{T}}. \tag{48}$$

Lemma 3. *Let $\chi : \mathbb{R}\times\mathbb{R}^3 \to \mathbb{R}^{3\times 3}$ be the susceptibility of the sample and $\vartheta \in S^2_+$ be given. Furthermore, let $R \in \mathrm{SO}(3)$ be such that there exists a constant $\alpha_R > 0$ and a direction $\vartheta_R \in S^2_+$ with*

$$\vartheta_R + e_3 = \alpha_R R(\vartheta + e_3) \tag{49}$$

and define the susceptibility χ_R of the rotated sample by (48).

Then, the data

$$\bar{\chi}_{R,p,j}(\tau;\sigma,\vartheta_R) = p_j \left[\vartheta_R \times \left(\vartheta_R \times \int_{E_{\sigma,\vartheta_R}} \chi_R(\tau, y)\mathrm{d}s(y)\right)\right]_j, \tag{50}$$

corresponding to the measurements of the rotated sample at the detector in direction ϑ_R, see (42), fulfil that

$$\bar{\chi}_{R,p,j}(\tau;\alpha_R\sigma,\vartheta_R) = p_j[\vartheta_R \times (\vartheta_R \times R\bar{\chi}(\tau;\sigma,\vartheta)R^{\mathrm{T}}]_j \tag{51}$$

for all $\tau,\sigma \in \mathbb{R}$, $p \in \mathbb{R}^2\times\{0\}$, $j = 1, 2$, where $\bar{\chi}$ is given by (43).

Proof. Inserting Definition (48) and substituting $z = R^{\mathrm{T}}y$, formula (50) becomes

$$\bar{\chi}_{R,p,j}(\tau;\sigma,\vartheta_R) = p_j \left[\vartheta_R \times \left(\vartheta_R \times \int_{R^{\mathrm{T}}E_{\sigma,\vartheta_R}} R\chi(\tau, z)R^{\mathrm{T}}\mathrm{d}s(z)\right)\right]_j.$$

Since now, by the Definition (38) of the plane $E_{\sigma,\vartheta}$ and by the Definition (49) of ϑ_R,

$$\begin{aligned} R^{\mathrm{T}}E_{\sigma,\vartheta_R} &= \{R^{\mathrm{T}}y \in \mathbb{R}^3 \mid \langle \vartheta_R + e_3, y\rangle = c\sigma\} \\ &= \{z \in \mathbb{R}^3 \mid \langle R^{\mathrm{T}}(\vartheta_R + e_3), z\rangle = c\sigma\} \\ &= \{z \in \mathbb{R}^3 \mid \alpha_R \langle \vartheta + e_3, z\rangle = c\sigma\} = E_{\frac{\sigma}{\alpha_R},\vartheta}, \end{aligned}$$

it follows (51). □

This means that the data $\bar{\chi}_{R,p,j}(\tau;\alpha_R\sigma,\vartheta_R)$ obtained from a detector placed in the direction ϑ_R, defined by (49), depends only on the Radon transform data $\bar{\chi}(\tau;\sigma,\vartheta)$. However, it still remains the algebraic problem of solving the Eq. (51) for different rotations R to obtain the matrix $\bar{\chi}(\tau;\sigma,\vartheta) \in \mathbb{R}^{3\times 3}$.

Proposition 12. *Let $A \in \mathbb{R}^{3\times 3}$ and $\vartheta \in S^2_+$ be given. Moreover, let $R_0, R_1, R_2 \in \mathrm{SO}(3)$ be rotations so that every proper subset of $\{R_0^{\mathrm{T}}e_3, R_1^{\mathrm{T}}e_3, R_2^{\mathrm{T}}e_3, \vartheta + e_3\}$ is linearly independent and such that there exist for every $R \in \{R_0, R_1, R_2\}$ constants $\alpha_R > 0$ and $\vartheta_R \in S^2_+$ fulfilling (49).*

Let further $P \in \mathbb{R}^{2\times 3}$ *be the orthogonal projection in direction* e_3, $P_\theta \in \mathbb{R}^{3\times 3}$ *the orthogonal projection in direction* $\theta \in \mathbb{R}^3$, *and*

$$B_R = \begin{pmatrix} -a_{R,e_1,1} & a_{R,e_1,1} - a_{R,e_1+e_2,1} \\ a_{R,e_2,2} - a_{R,e_1+e_2,2} & -a_{R,e_2,2} \end{pmatrix},$$
$$a_{R,p,j} = p_j[\vartheta_R \times (\vartheta_R \times RAR^{\mathrm{T}}p)]_j,$$

for every $R \in \{R_0, R_1, R_2\}$.

Then, the equations

$$PP_{\vartheta_R} RXR^{\mathrm{T}}P^{\mathrm{T}} = B_R, \quad R \in \{R_0, R_1, R_2\} \tag{52}$$

have the unique solution $X = A$.

Proof. Using that $\vartheta_R = \alpha_R R(\vartheta + e_3) - e_3$, see (49), it follows with $Pe_3 = 0$ that

$$PP_{\vartheta_R} = P(\mathbb{1} - \vartheta_R \vartheta_R^{\mathrm{T}}) = P - \alpha_R PR(\vartheta + e_3)\vartheta_R^{\mathrm{T}}.$$

With this identity, the Eq. (52) can be written in the form

$$PR(X - \alpha_R(\vartheta + e_3)\vartheta_R^{\mathrm{T}} RX)R^{\mathrm{T}}P^{\mathrm{T}} = B_R. \tag{53}$$

Let now $\eta_R \in \mathbb{R}^3$ denote a unit vector orthogonal to $\vartheta + e_3$ and orthogonal to $R^{\mathrm{T}}e_3$. Then $R\eta_R$ is orthogonal to e_3 and therefore, with $P^{\mathrm{T}}P = P_{e_3}$,

$$(PR\eta_R)^{\mathrm{T}} PR = (P_{e_3} R\eta_R)^{\mathrm{T}} R = (R\eta_R)^{\mathrm{T}} R = \eta_R^{\mathrm{T}}.$$

Thus, multiplying the Eq. (53) from the left with $(PR\eta_R)^{\mathrm{T}}$, it follows that

$$\eta_R^{\mathrm{T}}(X - \alpha_R(\vartheta + e_3)\vartheta_R^{\mathrm{T}} RX)R^{\mathrm{T}}P^{\mathrm{T}} = (PR\eta_R)^{\mathrm{T}} B_R.$$

Since now η_R is orthogonal to $\vartheta + e_3$, this simplifies to

$$\eta_R^{\mathrm{T}} XR^{\mathrm{T}}P^{\mathrm{T}} = (PR\eta_R)^{\mathrm{T}} B_R. \tag{54}$$

Remarking that the orthogonal projection onto the line $\mathbb{R}\eta_R$ can be written as the composition of two orthogonal projections onto orthogonal planes with intersection $\mathbb{R}\eta_R$:

$$\eta_R \eta_R^{\mathrm{T}} = P_{\nu_R} P_{\vartheta + e_3},$$

where ν_R is a unit vector orthogonal to η_R and $\vartheta + e_3$, and multiplying Eq. (54) from the left with η_R, one finds that

$$P_{\nu_R}(P_{\vartheta+e_3}X)R^{\mathrm{T}}P^{\mathrm{T}} = \eta_R(PR\eta_R)^{\mathrm{T}}B_R.$$

Applying then the projection P from the right and using that $R^{\mathrm{T}}P^{\mathrm{T}}P = R^{\mathrm{T}}P_{e_3} = P_{R^{\mathrm{T}}e_3}R^{\mathrm{T}}$, this can be written in the form

$$P_{\nu_R}(P_{\vartheta+e_3}X)P_{R^{\mathrm{T}}e_3} = \eta_R(PR\eta_R)^{\mathrm{T}}B_RPR. \tag{55}$$

Evaluating this equation now for $R = R_0, R_1, R_2$ and remarking that $\{\nu_{R_0}, \nu_{R_1}, \nu_{R_2}\}$ and $\{R_0^{\mathrm{T}}e_3, R_1^{\mathrm{T}}e_3, R_2^{\mathrm{T}}e_3\}$ are linearly independent, one concludes that the 3×3 matrix $P_{\vartheta+e_3}X$ is uniquely determined by (55).

However, it is also possible to calculate X (and not only its projection $P_{\vartheta+e_3}X$) from Eq. (53). Because

$$X = P_{\vartheta+e_3}X + \frac{1}{|\vartheta+e_3|^2}(\vartheta+e_3)(\vartheta+e_3)^{\mathrm{T}}X,$$

it follows from (53) that

$$PR\left(\frac{\vartheta+e_3}{|\vartheta+e_3|^2}(1-\alpha_R\vartheta_R^{\mathrm{T}}R(\vartheta+e_3))(\vartheta+e_3)^{\mathrm{T}}X\right)R^{\mathrm{T}}P^{\mathrm{T}}$$
$$= B_R - PR(\mathbb{1}-\alpha_R(\vartheta+e_3)\vartheta_R^{\mathrm{T}}R)P_{\vartheta+e_3}XR^{\mathrm{T}}P^{\mathrm{T}}. \tag{56}$$

Plugging in the identity

$$\alpha_R\vartheta_R^{\mathrm{T}}R(\vartheta+e_3) = \vartheta_R^{\mathrm{T}}(\vartheta_R+e_3) = 1+\vartheta_{R,3},$$

which follows from the Definition (49) of ϑ_R, applying P^{T} from the left and P from the right, and using as before $R^{\mathrm{T}}P^{\mathrm{T}}P = P_{R^{\mathrm{T}}e_3}R^{\mathrm{T}}$, the Eq. (56) yields

$$-\vartheta_{R,3}P_{R^{\mathrm{T}}e_3}\left(\frac{\vartheta+e_3}{|\vartheta+e_3|^2}(\vartheta+e_3)^{\mathrm{T}}X\right)P_{R^{\mathrm{T}}e_3}$$
$$= R^{\mathrm{T}}P^{\mathrm{T}}B_RPR - P_{R^{\mathrm{T}}e_3}(\mathbb{1}-\alpha_R(\vartheta+e_3)\vartheta_R^{\mathrm{T}})P_{\vartheta+e_3}XP_{R^{\mathrm{T}}e_3}. \tag{57}$$

Since the right hand side is already known (it depends only on $P_{\vartheta+e_3}X$), the equation system (57) for $R = R_0, R_1, R_2$ can be uniquely solved for

$$\frac{\vartheta+e_3}{|\vartheta+e_3|^2}(\vartheta+e_3)^{\mathrm{T}}X.$$

Therefore, the Eq. (52) uniquely determine X and because A is by construction a solution of the equations, this implies that $X = A$. □

Thus, applying Proposition 12 to the matrix $A = \bar{\chi}(\tau; \sigma, \vartheta)$ shows that the measurements $a_{R,p,j}$ obtained at the detectors ϑ_R for the polarizations $p = e_1, e_2, e_1 + e_2$ and rotations $R = R_0, R_1, R_2$, fulfilling the assumptions of Proposition 12 provide sufficient information to reconstruct the Radon data $\bar{\chi}(\tau; \sigma, \vartheta)$. Calculating these two-dimensional Radon data for all directions ϑ in some subset of S^2_+ (by considering some additional rotations so that for every direction ϑ, there exist three rotations fulfilling the assumptions of Proposition 12), it is possible via an inversion of a limited angle Radon transform to finally recover the susceptibility χ.

6 Conclusion

In this chapter, a general mathematical model of OCT based on Maxwell's equations has been presented. As a consequence of this modeling, OCT was formulated as an inverse scattering problem for the susceptibility χ. It was shown that without additional assumptions about the optical properties of the medium, in general, χ cannot be reconstructed due to lack of measurements. Some reasonable physical assumptions were presented, under which the medium can, in fact, be reconstructed. For instance, if the medium is isotropic, iterative schemes to reconstruct the susceptibility were developed. Dispersion and focus illumination are also considered. For an anisotropic medium, it follows that different incident fields, with respect to direction (rotating the sample) and polarization, should be considered to completely recover χ.

Acknowledgments The authors would like to thank Wolfgang Drexler and Boris Hermann from the Medical University Vienna for their valuable comments and stimulating discussions. This work has been supported by the Austrian Science Fund (FWF) within the national research network Photoacoustic Imaging in Biology and Medicine, projects S10501-N20 and S10505-N20.

Cross-References

▸ Inverse Scattering
▸ Optical Imaging
▸ Tomography
▸ Wave Phenomena

References

1. Ammari, H., Bao, G.: Analysis of the scattering map of a linearized inverse medium problem for electromagnetic waves. Inverse Prob. **17**, 219–234 (2001)
2. Andersen, P.E., Thrane, L., Yura, H.T., Tycho, A., Jørgensen, T.M., Frosz, M.H.: Advanced modelling of optical coherence tomography systems. Phys. Med. Biol. **49**, 1307–1327 (2004)
3. Born, M., Wolf, E.: Principles of Optics. 7th Edn. Cambridge University Press, Cambridge (1999)
4. Bouma, B.E., Tearney, G.J.: Handbook of Optical Coherence Tomography. Marcel Dekker Inc., New York (2002)

5. Brezinski, M.E.: Optical Coherence Tomography Principles and Applications. Academic Press, New York (2006)
6. Brodsky, A., Thurber, S.R., Burgess, L.W.: Low-coherence interferometry in random media. i. theory. J. Opt. Soc. Am. A **17**(11), 2024–2033 (2000)
7. Bruno, O., Chaubell, J.: One-dimensional inverse scattering problem for optical coherence tomography. Inverse Prob. **21**, 499–524 (2005)
8. Colton, D., Kress, R.: Inverse Acoustic and Electromagnetic Scattering Theory. 2nd edn, In: Applied Mathematical Sciences, vol. 93. Springer, Berlin (1998)
9. Dolin, L.S.: A theory of optical coherence tomography. Radiophys. Quantum Electron. **41**(10), 850–873 (1998)
10. Drexler, W., Fujimoto, J.G.: Optical Coherence Tomography. Springer, Berlin (2008)
11. Duan, L., Makita, S., Yamanari, M., Lim, Y., Yasuno, Y.: Monte-carlo-based phase retardation estimator for polarization sensitive optical coherence tomography. Opt. Express **19**, 16330–16345 (2011)
12. Feng, Y., Wang, R.K., Elder, J.B.: Theoretical model of optical coherence tomography for system optimization and characterization. J. Opt. Soc. Am. A **20**(9), 1792–1803 (2003)
13. Fercher, A.F.: Optical coherence tomography. J. Biomed. Opt. **1**(2), 157–173 (1996)
14. Fercher, A.F.: Optical coherence tomography - development, principles, applications. Z. Med. Phys. **20**, 251–276 (2010)
15. Fercher, A.F., Hitzenberger, C.K.: Optical Coherence Tomography. In: Progress in Optics. Elsevier Science B. V., Amsterdam (2002)
16. Fercher, A.F., Drexler, W., Hitzenberger, C.K., Lasser, T.: Optical coherence tomography - principles and applications. Rep. Prog. Phys. **66**(2), 239–303 (2003)
17. Fercher, A.F., Hitzenberger, C.K., Drexler, W., Kamp, G., Sattmann, H.: In vivo optical coherence tomography. Am. J. Ophthalmol. **116**, 113–114 (1993)
18. Fercher, A.F., Hitzenberger, C.K., Kamp, G., El Zaiat, S.Y.: Measurement of intraocular distances by backscattering spectral interferometry. Opt. Commun. **117**, 43–48 (1995)
19. Fercher, A.F., Sander, B., Jørgensen, T.M., Andersen, P.E.: Optical Coherence Tomography. In: Encyclopedia of Analytical Chemistry. John Wiley & Sons Ltd., Chichester (2009)
20. Friberg, A.T., Wolf, E.: Angular spectrum representation of scattered electromagnetic fields. J. Opt. Soc. Am. **73**(1), 26–32 (1983)
21. Hecht, E.: Optics. 4th edn. Addison Wesley, San Francisco (2002)
22. Hellmuth, T.: Contrast and resolution in optical coherence tomography. In: Bigio, I.J., Grundfest, W.S., Schneckenburger, H., Svanberg K., Viallet P.M., (eds.) Optical Biopsies and Microscopic Techniques. Proceedings of SPIE, vol 2926, pp 228–237 (1997)
23. Hohage, T.: Fast numerical solution of the electromagnetic medium scattering problem and applications to the inverse problem. J. Comput. Phys. **214**, 224–238 (2006)
24. Huang, D., Swanson, E.A., Lin, C.P., Schuman, J.S., Stinson, G., Chang, W., Hee, M.R., Flotte, T., Gregory, K., Puliafito, C.A., Fujimoto, J.G.: Optical coherence tomography. Science **254**(5035), 1178–1181 (1991)
25. Izatt, J.A., Choma, M.A.: Theory of optical coherence tomography. In: Drexler, W., Fujimoto, J.G. (eds.) In: Optical Coherence Tomography, pp. 47–72. Springer, Berlin (2008)
26. Kirillin, M., Meglinski, I., Kuzmin, V., Sergeeva, E., Myllylä, R.: Simulation of optical coherence tomography images by monte carlo modeling based on polarization vector approach. Opt. Express **18**(21), 21714–21724 (2010)
27. Knüttel, A., Schork, R., Böcker, D.: Analytical modeling of spatial resolution curves in turbid media acquired with optical coherence tomography (oct). In: Cogwell, C.J., Kino, G.S., Wilson, T. (eds.) Three- Dimensional Microscopy: Image Acquisition and Processing III, Proceedings of SPIE, vol 2655, pp. 258–270 (1996)
28. Marks, D.L., Davis, B.J., Boppart, S.A., Carney, P.S.: Partially coherent illumination in full-field interferometric synthetic aperture microscopy. J. Opt. Soc. Am. A **26**(2), 376–386 (2009)
29. Marks, D.L., Ralston, T.S., Boppart, S.A., Carney, P.S.: Inverse scattering for frequency-scanned full-field optical coherence tomography. J. Opt. Soc. Am. A **24**(4), 1034–1041 (2007)

30. Orfanidis, S.J.: Electromagnetic Waves and Antennas. Rutgers University Press, NJ (2002)
31. Pan, Y., Birngruber, R., Rosperich, J., Engelhardt, R.: Low-coherence optical tomography in turbid tissue: theoretical analysis. App. Opt. **34**(28), 6564–6574 (1995)
32. Podoleanu, A.G.: Optical coherence tomography. Br. J. Radiol. **78**, 976–988 (2005)
33. Potthast, R.: Integral equation methods in electromagnetic scattering from anisotropic media. Math. Methods Appl. Sci. **23**, 1145–1159 (2000)
34. Ralston, T.S.: Deconvolution methods for mitigation of transverse blurring in optical coherence tomography. IEEE Trans. Image Process. **14**(9), 1254–1264 (2005)
35. Ralston, T.S., Marks, D.L., Carney, P.S., Boppart, S.A.: Inverse scattering for optical coherence tomography. J. Opt. Soc. Am. A **23**(5), 1027–1037 (2006)
36. Schmitt, J.M.: Optical coherence tomography (OCT): A review. IEEE J. Quantum Electron. **5**, 1205–1215 (1999)
37. Schmitt, J.M., Knüttel, A.: Model of optical coherence tomography of heterogeneous tissue. J. Opt. Soc. Am. A **14**(6), 1231–1242 (1997)
38. Schmitt, J.M., Knüttel, A., Bonner, R.F.: Measurement of optical properties of biological tissues by low-coherence reflectometry. Appl. Opt. **32**, 6032–6042 (1993)
39. Schmitt, J.M., Xiang, S.H., Yung, K.M.: Differential absorption imaging with optical coherence tomography. J. Opt. Soc. Amer. A **15**, 2288–2296 (1998)
40. Smithies, D.J., Lindmo, T., Chen, Z., Nelson, J.S., Milner, T.E.: Signal attenuation and localization in optical coherence tomography studied by monte carlo simulation. Phys. Med. Biol. **43**, 3025–3044 (1998)
41. Swanson, E.A., Izatt, J.A., Hee, M.R., Huang, D., Lin, C.P., Schuman, J.S., Puliafito, C.A., Fujimoto, J.G.: In vivo retinal imaging by optical coherence tomography. Opt. Lett. **18**, 1864–1866 (1993)
42. Thomsen, J.B., Sander, B., Mogensen, M., Thrane, L., Jørgensen, T.M., Martini, T., Jemec, G.B.E., Andersen, P.E.: Optical coherence tomography: Technique and applications. In: Advanced Imaging in Biology and Medicine, pp. 103–129. Springer, Berlin (2009)
43. Thrane, L., Yura, H.T., Andersen, P.E.: Analysis of optical coherence tomography systems based on the extended huygens - fresnel principle. J. Opt. Soc. Am. A **17**(3), 484–490 (2000)
44. Tomlins, P.H., Wang, R.K.: Theory, developments and applications of optical coherence tomography. J. Phys. D: Appl. Phys. **38**, 2519–2535 (2005)
45. Turchin, I.V., Sergeeva, E.A., Dolin, L.S., Kamensky, V.A., Shakhova, N.M., Richards Kortum, R.: Novel algorithm of processing optical coherence tomography images for differentiation of biological tissue pathologies. J. Biomed. Opt. **10**(6) 064024, (2005)
46. Xu, C., Marks, D.L., Do, M.N., Boppart, S.A.: Separation of absorption and scattering profiles in spectroscopic optical coherence tomography using a least-squares algorithm. Opt. Express **12**(20), 4790–4803 (2004)

Wave Phenomena

Matti Lassas, Mikko Salo, and Gunther Uhlmann

Contents

1 Introduction 1206
2 Background 1207
Wave Imaging and Boundary Control Method 1207
Travel Times and Scattering Relation 1208
Curvelets and Wave Equations 1210
3 Mathematical Modeling and Analysis 1211
Boundary Control Method 1211
Travel Times and Scattering Relation 1234
Curvelets and Wave Equations 1239
4 Conclusion 1248
Cross-References 1248
References 1248

Abstract

This chapter discusses imaging methods related to wave phenomena, and in particular, inverse problems for the wave equation will be considered. The first part of the chapter explains the boundary control method for determining a wave speed of a medium from the response operator, which models boundary measurements. The second part discusses the scattering relation and travel times, which are different types of boundary data contained in the response operator. The third part gives a brief introduction to curvelets in wave imaging for media with nonsmooth wave speeds. The focus will be on theoretical results and methods.

M. Lassas (✉) • M. Salo
Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland
e-mail: matti.lassas@helsinki.fi

G. Uhlmann
Department of Mathematics, University of Washington, Seattle, WA, USA

© Springer Science+Business Media New York 2015

O. Scherzer (ed.), *Handbook of Mathematical Methods in Imaging*,
DOI 10.1007/978-1-4939-0790-8_52

1 Introduction

This chapter discusses imaging methods related to wave phenomena. Of the different types of waves that exist, we will focus on acoustic waves and problems which can be modeled by the acoustic wave equation. In the simplest case, this is the second-order linear hyperbolic equation

$$\partial_t^2 u(x,t) - c(x)^2 \Delta u(x,t) = 0$$

for a sound speed $c(x)$. This equation can be considered as a model for other hyperbolic equations, and the methods presented here can in some cases be extended to study wave phenomena in other fields such as electromagnetism or elasticity.

We will mostly be interested in inverse problems for the wave equation. In these problems, one has access to certain measurements of waves (the solutions u) on the surface of a medium, and one would like to determine material parameters (the sound speed c) of the interior of the medium from these boundary measurements. A typical field where such problems arise is seismic imaging, where one wishes to determine the interior structure of Earth by making various measurements of waves at the surface. We will not describe seismic imaging applications in more detail here, since they are discussed elsewhere in this volume.

Another feature in this chapter is that we will consistently consider *anisotropic* materials, where the sound speed depends on the direction of propagation. This means that the scalar sound speed $c(x)$, where $x = (x^1, x^2, \dots, x^n) \in \Omega \subset \mathbb{R}^n$, is replaced by a positive definite symmetric matrix $(g^{jk}(x))_{j,k=1}^n$, and the wave equation becomes

$$\partial_t^2 u(x,t) - \sum_{j,k=1}^n g^{jk}(x) \frac{\partial^2 u}{\partial x^j \partial x^k}(x,t) = 0.$$

Anisotropic materials appear frequently in applications such as in seismic imaging.

It will be convenient to interpret the anisotropic sound speed (g^{jk}) as the inverse of a Riemannian metric, thus modeling the medium as a *Riemannian manifold*. The benefits of such an approach are twofold. First, the well-established methods of Riemannian geometry become available to study the problems, and second, this provides an efficient way of dealing with the invariance under changes of coordinates present in many anisotropic wave imaging problems. The second point means that in inverse problems in anisotropic media, one can often only expect to recover the matrix (g^{jk}) up to a change of coordinates given by some diffeomorphism. In practice, this ambiguity could be removed by some a priori knowledge of the medium properties (such as the medium being in fact isotropic, see section “From Boundary Distance Functions to Riemannian Metric”).

2 Background

This chapter contains three parts which discuss different topics related to wave imaging. The first part considers the inverse problem of determining a sound speed in a wave equation from the response operator, also known as the hyperbolic Dirichlet-to-Neumann map, by using the boundary control method; see [5, 7, 42]. The second part considers other types of boundary measurements of waves, namely, the scattering relation and boundary distance function, and discusses corresponding inverse problems. The third part is somewhat different in nature and does not consider any inverse problems but rather gives an introduction to the use of curvelet decompositions in wave imaging for nonsmooth sound speeds. We briefly describe these three topics.

Wave Imaging and Boundary Control Method

Let us consider an isotropic wave equation. Let $\Omega \subset \mathbb{R}^n$ be an open, bounded set with smooth boundary $\partial\Omega$, and let $c(x)$ be a scalar-valued positive function in $C^\infty(\overline{\Omega})$ modeling the wave speed in Ω. First, we consider the wave equation

$$\begin{aligned} \partial_t^2 u(x,t) - c(x)^2 \Delta u(x,t) = 0 \quad &\text{in } \Omega \times \mathbb{R}_+, \\ u|_{t=0} = 0, \quad u_t|_{t=0} &= 0, \\ c(x)^{-n+1} \partial_{\mathbf{n}} u = f(x,t) \quad &\text{in } \partial\Omega \times \mathbb{R}_+, \end{aligned} \tag{1}$$

where $\partial_{\mathbf{n}}$ denotes the Euclidean normal derivative and n is the unit interior normal. We denote by $u^f = u^f(x,t)$ the solution of (1) corresponding to the boundary source term f.

Let us assume that the domain $\Omega \subset \mathbb{R}^n$ is known. The inverse problem is to reconstruct the wave speed $c(x)$ when we are given the set

$$\left\{ \left(f|_{\partial\Omega\times(0,2T)},\, u^f|_{\partial\Omega\times(0,2T)} \right) :\ f \in C_0^\infty(\partial\Omega \times \mathbb{R}_+) \right\},$$

that is, the Cauchy data of solutions corresponding to all possible boundary sources $f \in C_0^\infty(\partial\Omega \times \mathbb{R}_+)$, $T \in (0, \infty]$. If $T = \infty$, then this data is equivalent to the *response operator*

$$\Lambda_\Omega : f \mapsto u^f|_{\partial\Omega\times\mathbb{R}_+}, \tag{2}$$

which is also called the *nonstationary Neumann-to-Dirichlet map*. Physically, $\Lambda_\Omega f$ describes the measurement of the medium response to any applied boundary source f, and it is equivalent to various physical measurements. For instance, measuring how much energy is needed to force the boundary value $c(x)^{-n+1}\partial_{\mathbf{n}} u|_{\partial\Omega\times\mathbb{R}_+}$ to be equal to any given boundary value $f \in C_0^\infty(\partial\Omega \times \mathbb{R}_+)$ is equivalent to

measuring the map Λ_Ω on $\partial\Omega \times \mathbb{R}_+$; see [42,44]. Measuring Λ_Ω is also equivalent to measuring the corresponding Neumann-to-Dirichlet map for the heat or the Schrödinger equations or measuring the eigenvalues and the boundary values of the normalized eigenfunctions of the elliptic operator $-c(x)^2\Delta$; see [44].

The inverse problems for the wave equation and the equivalent inverse problems for the heat or the Schrödinger equations go back to works of M. Krein at the end of the 1950s, who used the causality principle in dealing with the one-dimensional inverse problem for an inhomogeneous string, $u_{tt} - c^2(x)u_{xx} = 0$; see, for example, [46]. In his works, causality was transformed into analyticity of the Fourier transform of the solution. A more straightforward hyperbolic version of the method was suggested by A. Blagovestchenskii at the end of 1960s to 1970s [12, 13]. The multidimensional case was studied by M. Belishev [4] in the late 1980s who understood the role of the PDE control for these problems and developed the boundary control method for hyperbolic inverse problems in domains of Euclidean space. Of crucial importance for the boundary control method was the result of D. Tataru in 1995 [77, 79] concerning a Holmgren-type uniqueness theorem for nonanalytic coefficients. The boundary control method was extended to the anisotropic case by M. Belishev and Y. Kurylev [7]. The geometric version of the boundary control method which we consider in this chapter was developed in [7,41,42,47]. We will consider the inverse problem in the more general setting of an anisotropic wave equation in an unbounded domain or on a non-compact manifold. These problems have been studied in detail in [39, 43] also in the case when the measurements are done only on a part of the boundary. In this paper we present a simplified construction method applicable for non-compact manifolds in the case when measurements are done on the whole boundary. We demonstrate these results in the case when we have an isotropic wave speed $c(x)$ in a bounded domain of Euclidean space. For this, we use the fact that in the Euclidean space, the only conformal deformation of a compact domain fixing the boundary is the identity map. This implies that after the abstract manifold structure (M, g) corresponding to the wave speed $c(x)$ in a given domain Ω is constructed, we can construct in an explicit way the embedding of the manifold M to the domain Ω and determine $c(x)$ at each point $x \in \Omega$. We note on the history of this result that using Tataru's unique continuation result [77], Theorem 2 concerning this case can be proven directly using the boundary control method developed for domains in Euclidean space in [4].

The reconstruction of non-compact manifolds has been considered also in [11, 27] with different kind of data, using iterated time reversal for solutions of the wave equation. We note that the boundary control method can be generalized also for Maxwell and Dirac equations under appropriate geometric conditions [50, 51], and its stability has been analyzed in [1, 45].

Travel Times and Scattering Relation

The problem considered in the previous section of recovering a sound speed from the response operator is highly overdetermined in dimensions $n \geq 2$. The Schwartz

kernel of the response operator depends on $2n$ variables, and the sound speed c depends on n variables.

In section "Travel Times and Scattering Relation," we will show that other types of boundary measurements in wave imaging can be directly obtained from the response operator. One such measurement is the *boundary distance function*, a function of $2n-2$ variables, which measures the travel times of shortest geodesics between boundary points. The problem of determining a sound speed from the travel times of shortest geodesics is the *inverse kinematic problem*. The more general problem of determining a Riemannian metric (corresponding to an anisotropic sound speed) up to isometry from the boundary distance function is the *boundary rigidity problem*. The problem is formally determined if $n = 2$ but overdetermined for $n \geq 3$.

This problem arose in geophysics in an attempt to determine the inner structure of the Earth by measuring the travel times of seismic waves. It goes back to Herglotz [37] and Wiechert and Zoeppritz [84] who considered the case of a radial metric conformal to the Euclidean metric. Although the emphasis has been in the case that the medium is isotropic, the anisotropic case has been of interest in geophysics since the Earth is anisotropic. It has been found that even the inner core of the Earth exhibits anisotropic behavior [24].

To give a proper definition of the boundary distance function, we will consider a bounded domain $\Omega \subset \mathbb{R}^n$ with smooth boundary to be equipped with a Riemannian metric g, that is, a family of positive definite symmetric matrices $g(x) = \big(g_{jk}(x)\big)_{j,k=1}^n$ depending smoothly on $x \in \overline{\Omega}$. The length of a smooth curve $\gamma : [a,b] \to \overline{\Omega}$ is defined to be

$$L_g(\gamma) = \int_a^b \left(\sum_{j,k=1}^n g_{jk}\big(\gamma(t)\big)\dot{\gamma}^j(t)\dot{\gamma}^k(t) \right)^{1/2} dt.$$

The distance function $d_g(x,y)$ for $x, y \in \overline{\Omega}$ is the infimum of the lengths of all piecewise smooth curves in $\overline{\Omega}$ joining x and y. The boundary distance function is $d_g(x,y)$ for $x, y \in \partial\Omega$.

In the boundary rigidity problem, one would like to determine a Riemannian metric g from the boundary distance function d_g. In fact, since $d_g = d_{\psi^* g}$ for any diffeomorphism $\psi : \overline{\Omega} \to \overline{\Omega}$ which fixes each boundary point, we are looking to recover from d_g the metric g up to such a diffeomorphism. Here, $\psi^* g(y) = D\psi(y)^t g\big(\psi(y)\big) D\psi(y)$ is the pullback of g by ψ.

It is easy to give counterexamples showing that this cannot be done in general; consider, for instance, the closed hemisphere, where boundary distances are given by boundary arcs so making the metric larger in the interior does not change d_g. Michel [55] conjectured that a *simple* metric g is uniquely determined, up to an action of a diffeomorphism fixing the boundary, by the boundary distance function $d_g(x,y)$ known for all x and y on $\partial\Omega$. A metric is called simple if for any two

points in $\overline{\Omega}$, there is a unique length minimizing geodesic joining them, and if the boundary is strictly convex.

The conjecture of Michel has been proved for two-dimensional simple manifolds [60]. In higher dimensions, it is open, but several partial results are known, including the recent results of Burago and Ivanov for metrics close to Euclidean [15] and close to hyperbolic [16] (see the survey [40]). Earlier and related works include results for simple metrics conformal to each other [8, 10, 26, 56–58], for flat metrics [34], for locally symmetric spaces of negative curvature [9], for two-dimensional simple metrics with negative curvature [25, 59], a local result [70], a semiglobal solvability result [54], and a result for generic simple metrics [71].

In case the metric is not simple, instead of the boundary distance function, one can consider the more general *scattering relation* which encodes, for any geodesic starting and ending at the boundary, the start point and direction, the end point and direction, and the length of the geodesic. We will see in section "Travel Times and Scattering Relation" that also this information can be determined directly from the response operator. If the metric is simple, then the scattering relation and boundary distance function are equivalent, and either one is determined by the other.

The *lens rigidity problem* is to determine a metric up to isometry from the scattering relation. There are counterexamples of manifolds which are trapping, and the conjecture is that on a nontrapping manifold the metric is determined by the scattering relation up to isometry. We refer to [72] and the references therein for known results on this problem.

Curvelets and Wave Equations

In section "Curvelets and Wave Equations," we describe an alternative approach to the analysis of solutions of wave equations, based on a decomposition of functions into basic elements called *curvelets* or *wave packets*. This approach also works for wave speeds of limited smoothness unlike some of the approaches presented earlier. Furthermore, the curvelet decomposition yields efficient representations of functions containing sharp wave fronts along curves or surfaces, thus providing a common framework for representing such data and analyzing wave phenomena and imaging operators. Curvelets and related methods have been proposed as computational tools for wave imaging, and the numerical aspects of the theory are a subject of ongoing research.

A curvelet decomposition was introduced by Smith [67] to construct a solution operator for the wave equation with $C^{1,1}$ sound speed and to prove Strichartz estimates for such equations. This started a body of research on L^p estimates for low-regularity wave equations based on curvelet-type methods; see, for instance, Tataru [80–82], Smith [68], and Smith and Sogge [69]. Curvelet decompositions have their roots in harmonic analysis and the theory of Fourier integral operators,

where relevant works include Córdoba and Fefferman [23] and Seeger et al. [65] (see also Stein [73]).

In a rather different direction, curvelet decompositions came up in image analysis as an optimally sparse way of representing images with C^2 edges; see Candés and Donoho [20] (the name "curvelet" was introduced in [19]). The property that curvelets yield sparse representations for wave propagators was studied in Candés and Demanet [17, 18]. Numerical aspects of curvelet-type methods in wave computation are discussed in [21, 30]. Finally, both theoretical and practical aspects of curvelet methods related to certain seismic imaging applications are studied in [2, 14, 29, 31, 64].

3 Mathematical Modeling and Analysis

Boundary Control Method

Inverse Problems on Riemannian Manifolds

Let $\Omega \subset \mathbb{R}^n$ be an open, bounded set with smooth boundary $\partial\Omega$ and let $c(x)$ be a scalar-valued positive function in $C^\infty(\overline{\Omega})$, modeling the wave speed in Ω. We consider the closure $\overline{\Omega}$ as a differentiable manifold M with a smooth, nonempty boundary. We consider also a more general case, and allow (M, g) to be a possibly non-compact, complete manifold with boundary. This means that the manifold contains its boundary ∂M and M is complete with metric d_g defined below. Moreover, near each point $x \in M$, there are coordinates (U, X), where $U \subset M$ is a neighborhood of x and $X : U \to \mathbb{R}^n$ if x is an interior point, or $X : U \to \mathbb{R}^{n-1} \times [0, \infty)$ if x is a boundary point such that for any coordinate neighborhoods (U, X) and $(\tilde{U}, \tilde{X})$, the transition functions $X \circ \tilde{X}^{-1} : \tilde{X}(U \cap \tilde{U}) \to X(U \cap \tilde{U})$ are C^∞-smooth. Note that all compact Riemannian manifolds are complete according to this definition. Usually we denote the components of X by $X(y) = \big(x^1(y), \ldots, x^n(y)\big)$.

Let u be the solution of the wave equation

$$\begin{aligned} u_{tt}(x,t) + Au(x,t) &= 0 \quad \text{in} \quad M \times \mathbb{R}_+, \\ u|_{t=0} &= 0, \quad u_t|_{t=0} = 0, \\ B_{\nu,\eta} u|_{\partial M \times \mathbb{R}_+} &= f. \end{aligned} \tag{3}$$

Here, $f \in C_0^\infty(\partial M \times \mathbb{R}_+)$ is a real-valued function, and $A = A(x, D)$ is an elliptic partial differential operator of the form

$$Av = -\sum_{j,k=1}^{n} \mu(x)^{-1} |g(x)|^{-\frac{1}{2}} \frac{\partial}{\partial x^j} \left(\mu(x) |g(x)|^{\frac{1}{2}} g^{jk}(x) \frac{\partial v}{\partial x^k}(x) \right) + q(x) v(x), \tag{4}$$

where $g^{jk}(x)$ is a smooth, symmetric, real, positive definite matrix, $|g| = \det\big(g^{jk}(x)\big)^{-1}$, and $\mu(x) > 0$ and $q(x)$ are smooth real-valued functions. On

existence and properties of the solutions of Eq. (3), see [52]. The inverse of the matrix $\big(g^{jk}(x)\big)_{j,k=1}^n$, denoted $\big(g_{jk}(x)\big)_{j,k=1}^n$ defines a Riemannian metric on M. The tangent space of M at x is denoted by T_xM, and it consists of vectors p which in local coordinates (U, X), $X(y) = \big(x^1(y), \dots, x^n(y)\big)$ are written as $p = \sum_{k=1}^n p^k \frac{\partial}{\partial x^k}$. Similarly, the cotangent space T_x^*M of M at x consists of covectors which are written in the local coordinates as $\xi = \sum_{k=1}^n \xi_k dx^k$. The inner product which g determines in the cotangent space T_x^*M of M at the point x is denoted by $\langle \xi, \eta \rangle_g = g(\xi, \eta) = \sum_{j,k=1}^n g^{jk}(x)\xi_j \eta_k$ for $\xi, \eta \in T_x^*M$. We use the same notation for the inner product at the tangent space T_xM, that is, $\langle p, q \rangle_g = g(p, q) = \sum_{j,k=1}^n g_{jk}(x) p^j q^k$ for $p, q \in T_xM$.

The metric defines a distance function, which we call also the travel time function,

$$d_g(x, y) = \inf |\mu|, \quad |\mu| = \int_0^1 \langle \partial_s \mu(s), \partial_s \mu(s) \rangle_g^{1/2} ds,$$

where $|\mu|$ denotes the length of the path μ, and the infimum is taken over all piecewise C^1-smooth paths $\mu : [0, 1] \to M$ with $\mu(0) = x$ and $\mu(1) = y$.

We define the space $L^2(M, dV_\mu)$ with inner product

$$\langle u, v \rangle_{L^2(M, dV_\mu)} = \int_M u(x) v(x)\, dV_\mu(x),$$

where $dV_\mu = \mu(x)|g(x)|^{1/2} dx^1 dx^2 \dots dx^n$. By the above assumptions, A is formally self-adjoint, that is,

$$\langle Au, v \rangle_{L^2(M, dV_\mu)} = \langle u, Av \rangle_{L^2(M, dV_\mu)} \quad \text{for } u, v \in C_0^\infty(M^{\text{int}}).$$

Furthermore, let

$$B_{\nu,\eta} v = -\partial_\nu v + \eta v,$$

where $\eta : \partial M \to \mathbb{R}$ is a smooth function and

$$\partial_\nu v = \sum_{j,k=1}^n \mu(x) g^{jk}(x) \nu_k \frac{\partial}{\partial x^j} v(x),$$

where $\nu(x) = (\nu_1, \nu_2, \dots, \nu_m)$ is the interior conormal vector field of ∂M, satisfying $\sum_{j,k=1}^n g^{jk} \nu_j \xi_k = 0$ for all cotangent vectors of the boundary, $\xi \in T^*(\partial M)$. We assume that ν is normalized, so that $\sum_{j,k=1}^n g^{jk} \nu_j \nu_k = 1$. If M is compact, then the operator A in the domain $\mathcal{D}(A) = \{v \in H^2(M) : \partial_\nu v|_{\partial M} = 0\}$, where $H^s(M)$ denotes the Sobolev spaces on M, is an unbounded self-adjoint operator in $L^2(M, dV_\mu)$.

An important example is the operator

$$A_0 = -c^2(x)\Delta + q(x) \tag{5}$$

on a bounded smooth domain $\Omega \subset \mathbb{R}^n$ with $\partial_\nu v = c(x)^{-n+1}\partial_{\mathbf{n}} v$, where $\partial_{\mathbf{n}} v$ is the Euclidean normal derivative of v.

We denote the solutions of (3) by

$$u(x,t) = u^f(x,t).$$

For the initial boundary value problem (3), we define the nonstationary Robin-to-Dirichlet map or the response operator Λ by

$$\Lambda f = u^f|_{\partial M \times \mathbb{R}_+}. \tag{6}$$

The finite time response operator Λ^T corresponding to the finite observation time $T > 0$ is given by

$$\Lambda^T f = u^f|_{\partial M \times (0,T)}. \tag{7}$$

For any set $B \subset \partial M \times \mathbb{R}_+$, we denote $L^2(B) = \{f \in L^2(\partial M \times \mathbb{R}_+) : \operatorname{supp}(f) \subset B\}$. This means that we identify the functions and their zero continuations.

By [78], the map Λ^T can be extended to bounded linear map $\Lambda^T : L^2(B) \to H^{1/3}\big(\partial M \times (0,T)\big)$ when $B \subset \partial M \times (0,T)$ is compact. Here, $H^s\big(\partial M \times (0,T)\big)$ denotes the Sobolev space on $\partial M \times (0,T)$. Below we consider Λ^T also as a linear operator $\Lambda^T : L^2_{cpt}\big(\partial M \times (0,T)\big) \to L^2\big(\partial M \times (0,T)\big)$, where $L^2_{cpt}\big(\partial M \times (0,T)\big)$ denotes the compactly supported functions in $L^2\big(\partial M \times (0,T)\big)$.

For $t > 0$ and a relatively compact open set $\Gamma \subset \partial M$, let

$$M(\Gamma, t) = \{x \in M \ : \ d_g(x,\Gamma) < t\}. \tag{8}$$

This set is called the domain of influence of Γ at time t.

When $\Gamma \subset \partial M$ is an open relatively compact set and $f \in C_0^\infty(\Gamma \times \mathbb{R}_+)$, it follows from finite speed of wave propagation (see, e.g., [38]) that the wave $u^f(t) = u^f(\cdot, t)$ is supported in the domain $M(\Gamma, t)$, that is,

$$u^f(t) \in L^2\big(M(\Gamma,t)\big) = \{v \in L^2(M) : \ \operatorname{supp}(v) \subset M(\Gamma,t)\}. \tag{9}$$

We will consider the boundary of the manifold ∂M with the metric $g_{\partial M} = \iota^* g$ inherited from the embedding $\iota : \partial M \to M$. We assume that we are given the *boundary data*, that is, the collection

$$(\partial M, g_{\partial M}) \text{ and } \Lambda, \tag{10}$$

where $(\partial M, g_{\partial M})$ is considered as a smooth Riemannian manifold with a known differentiable and metric structure and Λ is the nonstationary Robin-to-Dirichlet map given in (6).

Our goal is to reconstruct the isometry type of the Riemannian manifold (M, g), that is, a Riemannian manifold which is isometric to the manifold (M, g). This is often stated by saying that we reconstruct (M, g) up to an isometry. Our next goal is to prove the following result:

Theorem 1. *Let (M, g) to be a smooth, complete Riemannian manifold with a nonempty boundary. Assume that we are given the boundary data (10). Then it is possible to determine the isometry type of manifold (M, g).*

From Boundary Distance Functions to Riemannian Metric

In order to reconstruct (M, g), we use a special representation, the *boundary distance representation*, $R(M)$, of M and later show that the boundary data (10) determine $R(M)$. We consider next the (possibly unbounded) continuous functions $h : C(\partial M) \to \mathbb{R}$. Let us choose a specific point $Q_0 \in \partial M$ and a constant $C_0 > 0$ and using these, endow $C(\partial M)$ with the metric

$$d_C(h_1, h_2) = |h_1(Q_0) - h_2(Q_0)| + \sup_{z \in \partial M} \min(C_0, |h_1(z) - h_2(z)|). \tag{11}$$

Consider a map $R : M \to C(\partial M)$,

$$R(x) = r_x(\cdot); \quad r_x(z) = d_g(x, z), \ z \in \partial M, \tag{12}$$

that is, $r_x(\cdot)$ is the *distance function* from $x \in M$ to the points on ∂M. The image $R(M) \subset C(\partial M)$ of R is called the boundary distance representation of M. The set $R(M)$ is a metric space with the distance inherited from $C(\partial M)$ which we denote by d_C, too. The map R, due to the triangular inequality, is Lipschitz,

$$d_C(r_x, r_y) \leq 2 d_g(x, y). \tag{13}$$

We note that when M is compact and $C_0 = \operatorname{diam}(M)$, the metric $d_C : C(\partial M) \to \mathbb{R}$ is a norm which is equivalent to the standard norm $\|f\|_\infty = \max_{x \in \partial M} |f(x)|$ of $C(\partial M)$.

We will see below that the map $R : M \to R(M) \subset C(\partial M)$ is an embedding. Many results of differential geometry, such as Whitney or Nash embedding theorems, concern the question how an abstract manifold can be embedded to some simple space such as a higher dimensional Euclidean space. In the inverse problem, we need to construct a "copy" of the unknown manifold in some known space, and as we assume that the boundary is given, we do this by embedding the manifold M to the known, although infinite dimensional function space $C(\partial M)$.

Next we recall some basic definitions on Riemannian manifolds; see, for example, [22] for an extensive treatment. A path $\mu : [a, b] \to N$ is called a

geodesic if, for any $c \in [a,b]$, there is $\varepsilon > 0$ such that if $s,t \in [a,b]$ such that $c-\varepsilon < s < t < c+\varepsilon$, the path $\mu([s,t])$ is a shortest path between its endpoints, that is,

$$|\mu([s,t])| = d_g\big(\mu(s),\mu(t)\big).$$

In the future, we will denote a geodesic path μ by γ and parameterize γ with its arclength s, so that $|\mu([s_1,s_2])| = d_g\big(\mu(s_1),\mu(s_2)\big)$. Let $x(s)$,

$$x(s) = \big(x^1(s),\dots,x^n(s)\big),$$

be the representation of the geodesic γ in local coordinates (U,X). In the interior of the manifold, that is, for $U \subset M^{\text{int}}$ the path $x(s)$ satisfies the second-order differential equations

$$\frac{d^2x^k(s)}{ds^2} = -\sum_{i,j=1}^{n} \Gamma_{ij}^k\big(x(s)\big)\frac{dx^i(s)}{ds}\frac{dx^j(s)}{ds}, \tag{14}$$

where Γ_{ij}^k are the Christoffel symbols, given in local coordinates by the formula

$$\Gamma_{ij}^k(x) = \sum_{p=1}^{n} \frac{1}{2} g^{kp}(x)\left(\frac{\partial g_{jp}}{\partial x^i}(x) + \frac{\partial g_{ip}}{\partial x^j}(x) - \frac{\partial g_{ij}}{\partial x^p}(x)\right).$$

Let $y \in M$ and $\xi \in T_xM$ be a unit vector satisfying the condition $g\big(\xi,\nu(y)\big) > 0$ in the case when $y \in \partial M$. Then, we can consider the solution of the initial value problem for the differential equation (14) with the initial data

$$x(0) = y, \quad \frac{dx}{ds}(0) = \xi.$$

This initial value problem has a unique solution $x(s)$ on an interval $[0, s_0(y,\xi))$ such that $s_0(y,\xi) > 0$ is the smallest value $s_0 > 0$ for which $x(s_0) \in \partial M$, or $s_0(y,\xi) = \infty$ in case no such s_0 exists. We will denote $x(s) = \gamma_{y,\xi}(s)$ and say that the geodesic is a normal geodesic starting at y if $y \in \partial M$ and $\xi = \nu(y)$.

Example 1. In the case when (M,g) is such a compact manifold that all geodesics are the shortest curves between their endpoints and all geodesics can be continued to geodesics that hit the boundary, we can see that the metric spaces (M,d_g) and $(R(M), \|\cdot\|_\infty)$ are isometric. Indeed, for any two points $x,y \in M$, there is a geodesic γ from x to a boundary point z, which is a continuation of the geodesic from x to y. As in the considered case the geodesics are distance minimizing curves, we see that

$$r_x(z) - r_y(z) = d_g(x,z) - d_g(y,z) = d_g(x,y),$$

and thus $\|r_x - r_y\|_\infty \geq d_g(x,y)$. Combining this with the triangular inequality, we see that $\|r_x - r_y\|_\infty = d_g(x,y)$ for $x, y \in M$ and R is isometry of (M, d_g) and $(R(M), \|\cdot\|_\infty)$.

Notice that when even M is a compact manifold, the metric spaces (M, d_g) and $(R(M), \|\cdot\|_\infty)$ are not always isometric. As an example, consider a unit sphere in $\mathbb{R}^3$ with a small circular hole near the South pole of, say, diameter ε. Then, for any x, y on the equator and $z \in \partial M$, $\pi/2 - \varepsilon \leq r_x(z) \leq \pi/2$ and $\pi/2 - \varepsilon \leq r_y(z) \leq \pi/2$. Then $d_C(r_x, r_y) \leq \varepsilon$, while $d_g(x,y)$ may be equal to π.

Next, we introduce the *boundary normal coordinates* on M. For a normal geodesic $\gamma_{z,v}(s)$ starting from $z \in \partial M$ consider $d_g(\gamma_{z,v}(s), \partial M)$. For small s,

$$d_g(\gamma_{z,v}(s), \partial M) = s, \tag{15}$$

and z is the unique nearest point to $\gamma_{z,v}(s)$ on ∂M. Let $\tau(z) \in (0, \infty]$ be the largest value for which (15) is valid for all $s \in [0, \tau(z)]$. Then for $s > \tau(z)$,

$$d_g(\gamma_{z,v}(s), \partial M) < s,$$

and z is no more the nearest boundary point for $\gamma_{z,v}(s)$. The function $\tau(z) \in C(\partial M)$ is called the cut locus distance function, and the set

$$\omega = \{\gamma_{z,v}\big(\tau(z)\big) \in M : z \in \partial M, \text{ and } \tau(z) < \infty\}, \tag{16}$$

is the *cut locus of M with respect to ∂M*. The set ω is a closed subset of M having zero measure. In particular, $M \backslash \omega$ is dense in M. In the remaining domain $M \backslash \omega$, we can use the coordinates

$$x \mapsto \big(z(x), t(x)\big), \tag{17}$$

where $z(x) \in \partial M$ is the unique nearest point to x and $t(x) = d_g(x, \partial M)$. (Strictly speaking, one also has to use some local coordinates of the boundary, $y : z \mapsto \big(y^1(z), \ldots, y^{(n-1)}(z)\big)$ and define that

$$x \mapsto \big(y\big(z(x)\big), t(x)\big) = \big(y^1(z(x)\big), \ldots, y^{(n-1)}\big(z(x)\big), t(x)\big) \in \mathbb{R}^n, \tag{18}$$

are the boundary normal coordinates.) Using these coordinates, we show that $R : M \to C(\partial M)$ is an embedding. The result of Lemma 1 is considered in detail for compact manifolds in [42].

Lemma 1. *Let (M, d_g) be the metric space corresponding to a complete Riemannian manifold (M, g) with a nonempty boundary. The map $R : (M, d_g) \to (R(M), d_C)$ is a homeomorphism. Moreover, given $R(M)$ as a subset of $C(\partial M)$, it is possible to construct a distance function d_R on $R(M)$ that makes the metric space $(R(M), d_R)$ isometric to (M, d_g).*

Proof. We start by proving that R is a homeomorphism. Recall the following simple result from topology:

Assume that X and Y are Hausdorff spaces, X is compact, and $F : X \to Y$ is a continuous, bijective map from X to Y. Then $F : X \to Y$ is a homeomorphism.

Let us next extend this principle. Assume that (X, d_X) and (Y, d_Y) are metric spaces and let $X_j \subset X$, $j \in \mathbb{Z}_+$ be compact sets such that $\bigcup_{j\in\mathbb{Z}_+} X_j = X$. Assume that $F : X \to Y$ is a continuous, bijective map. Moreover, let $Y_j = F(X_j)$ and assume that there is a point $p \in Y$ such that

$$a_j = \inf_{y\in Y\setminus Y_j} d_Y(y, p) \to \infty \text{ as } j \to \infty. \tag{19}$$

Then by the above, the maps $F : \cup_{j=1}^n X_j \to \cup_{j=1}^n Y_j$ are homeomorphisms for all $n \in \mathbb{Z}_+$. Next, consider a sequence $y_k \in Y$ such that $y_k \to y$ in Y as $k \to \infty$. By removing first elements of the sequence $(y_k)_{k=1}^\infty$ if needed, we can assume that $d_Y(y_k, y) \le 1$. Let now $N \in \mathbb{Z}_+$ be such that for $j > N$, we have $a_j > b := d_Y(y, p) + 1$. Then $y_k \in \bigcup_{j=1}^N Y_j$ and as the map $F : \bigcup_{j=1}^N X_j \to \bigcup_{j=1}^N Y_j$ is a homeomorphism, we see that $F^{-1}(y_k) \to F^{-1}(y)$ in X as $k \to \infty$. This shows that $F^{-1} : Y \to X$ is continuous, and thus $F : X \to Y$ is a homeomorphism.

By definition, $R : M \to R(M)$ is surjective and, by (13), continuous. In order to prove the injectivity, assume the contrary, that is, $r_x(\cdot) = r_y(\cdot)$ but $x \neq y$. Denote by z_0 any point where

$$\min_{z\in\partial M} r_x(z) = r_x(z_0).$$

Then

$$\begin{aligned} d_g(x, \partial M) &= \min_{z\in\partial M} r_x(z) = r_x(z_0) \\ &= r_y(z_0) = \min_{z\in\partial M} r_y(z) = d_g(y, \partial M), \end{aligned} \tag{20}$$

and $z_0 \in \partial M$ is a nearest boundary point to x. Let μ_x be the shortest path from z_0 to x. Then, the path μ_x is a geodesic from x to z_0 which intersects ∂M first time at z_0. By using the first variation on length formula, we see that μ_x has to hit to z_0 normally; see [22]. The same considerations are true for the point y with the same point z_0. Thus, both x and y lie on the normal geodesic $\gamma_{z_0,\nu}(s)$ to ∂M. As the geodesics are unique solutions of a system of ordinary differential equations (the Hamilton–Jacobi equation (14)), they are uniquely determined by their initial points

and directions, that is, the geodesics are non-branching. Thus, we see that

$$x = \gamma_{z_0}(s_0) = y,$$

where $s_0 = r_x(z_0) = r_y(z_0)$. Hence, $R : M \to C(\partial M)$ is injective.

Next, we consider the condition (19) for $R : M \to R(M)$. Let $z \in M$ and consider closed sets $X_j = \{x \in M : d_C\big(R(x), R(z)\big) \leq j\}$, $j \in \mathbb{Z}_+$. Then for $x \in X_j$, we have by definition (11) of the metric d_C that

$$d_g(x, Q_0) \leq j + d_g(z, Q_0),$$

implying that the sets X_j, $j \in \mathbb{Z}_+$ are compact. Clearly, $\bigcup_{j\in\mathbb{Z}_+} X_j = X$. Let next $Y_j = R(X_j) \subset Y = R(M)$ and $p = R(Q_0) \in R(M)$. Then for $r_x \in Y \setminus Y_j$, we have

$$\begin{aligned} d_C(r_x, p) &\geq r_x(Q_0) - p(Q_0) = d_g(x, Q_0) \\ &\geq j - d_g(z, Q_0) - C_0 \to \infty \text{ as } j \to \infty \end{aligned}$$

and thus the condition (19) is satisfied. As $R : M \to R(M)$ is a continuous, bijective map, it implies that $R : M \to R(M)$ is a homeomorphism.

Next we introduce a differentiable structure and a metric tensor, g_R, on $R(M)$ to have an isometric diffeomorphism

$$R : (M, g) \to (R(M), g_R). \tag{21}$$

Such structures clearly exists – the map R pushes the differentiable structure of M and the metric g to some differentiable structure on $R(M)$ and the metric $g_R := R_* g$ which makes the map (21) an isometric diffeomorphism. Next we construct these coordinates and the metric tensor in those on $R(M)$ using the fact that $R(M)$ is known as a subset of $C(\partial M)$.

We will start by construction of the differentiable and metric structures on $R(M)\setminus R(\omega)$, where ω is the cut locus of M with respect to ∂M. First, we show that we can identify in the set $R(M)$ all the elements of the form $r = r_x \in R(M)$ where $x \in M \setminus \omega$. To do this, we observe that $r = r_x$ with $x = \gamma_{z,v}(s)$, $s < \tau(z)$ if and only if:

1. $r(\cdot)$ has a unique global minimum at some point $z \in \partial M$;
2. There is $\tilde{r} \in R(M)$ having a unique global minimum at the same z and $r(z) < \tilde{r}(z)$. This is equivalent to saying that there is y with $r_y(\cdot)$ having a unique global minimum at the same z and $r_x(z) < r_y(z)$.

Thus, we can find $R(M \setminus \omega)$ by choosing all those $r \in R(M)$ for which the above conditions (1) and (2) are valid.

Next, we choose a differentiable structure on $R(M\backslash\omega)$ which makes the map $R : M\backslash\omega \to R(M\backslash\omega)$ a diffeomorphism. This can be done by introducing coordinates near each $r^0 \in R(M\backslash\omega)$. In a sufficiently small neighborhood $W \subset R(M)$ of r^0 the coordinates

$$r \mapsto \big(Y(r), T(r)\big) = \left(y(\operatorname{argmin}_{z\in\partial M} r), \min_{z\in\partial M} r\right)$$

are well defined. These coordinates have the property that the map $x \mapsto \big(Y(r_x), T(r_x)\big)$ coincides with the boundary normal coordinates (17) and (18). When we choose the differential structure on $R(M\backslash\omega)$ that corresponds to these coordinates, the map

$$R : M\backslash\omega \to R(M\backslash\omega)$$

is a diffeomorphism.

Next we construct the metric g_R on $R(M)$. Let $r^0 \in R(M\backslash\omega)$. As above, in a sufficiently small neighborhood $W \subset R(M)$ of r^0, there are coordinates $r \mapsto X(r) := \big(Y(r), T(r)\big)$ that correspond to the boundary normal coordinates. Let $(y^0, t^0) = X(r^0)$. We consider next the evaluation function

$$K_w : W \to \mathbb{R}, \quad K_w(r) = r(w),$$

where $w \in \partial M$. The inverse of $X : W \to \mathbb{R}^n$ is well defined in a neighborhood $U \subset \mathbb{R}^n$ of (y^0, t^0), and thus we can define the function

$$E_w = K_w \circ X^{-1} : U \to \mathbb{R}$$

that satisfies

$$E_w(y,t) := d_g\left(w, \gamma_{z(y),\nu(y)}(t)\right), \quad (y,t) \in U, \tag{22}$$

where $\gamma_{z(y),\nu(y)}(t)$ is the normal geodesic starting from the boundary point $z(y)$ with coordinates $y = (y^1, \dots, y^{n-1})$ and $\nu(y)$ is the interior unit normal vector at y.

Let now $g_R = R_* g$ be the push-forward of g to $R(M\backslash\omega)$. We denote its representation in X-coordinates by $g_{jk}(y,t)$. Since X corresponds to the boundary normal coordinates, the metric tensor satisfies

$$g_{mm} = 1, \quad g_{\alpha m} = 0, \quad \alpha = 1, \dots, n-1.$$

Consider the function $E_w(y,t)$ as a function of (y,t) with a fixed w. Then its differential, dE_w at point (y,t) defines a covector in $T^*_{(y,t)}(U) = \mathbb{R}^n$. Since the gradient of a distance function is a unit vector field, we see from (22) that

$$\|dE_w(y,t)\|^2_{(g_{jk})} := \left(\frac{\partial}{\partial t}E_w(y,t)\right)^2 + \sum_{\alpha,\beta=1}^{n-1}(g_R)^{\alpha\beta}(y,t)\frac{\partial E_w}{\partial y^\alpha}(y,t)\frac{\partial E_w}{\partial y^\beta}(y,t) = 1.$$

Let us next fix a point $(y^0,t^0) \in U$. Varying the point $w \in \partial M$, we obtain a set of covectors $dE_w(y^0,t^0)$ in the unit ball of $\left(T^*_{(y^0,t^0)}U, g_{jk}\right)$ which contains an open neighborhood of $(0,\dots,0,1)$. This determines uniquely the tensor $g^{jk}(y^0,t^0)$. Thus, we can construct the metric tensor in the boundary normal coordinates at arbitrary $r \in R(M\backslash\omega)$. This means that we can find the metric g_R on $R(M\backslash\omega)$ when $R(M)$ is given.

To complete the reconstruction, we need to find the differentiable structure and the metric tensor near $R(\omega)$. Let $r^{(0)} \in R(\omega)$ and $x^{(0)} \in M^{\text{int}}$ be such a point that $r^{(0)} = r_{x^{(0)}} = R(x^{(0)})$. Let z_0 be some of the closest points of ∂M to the point $x^{(0)}$. Then there are points $z_1,\dots,z_{n-1}$ on ∂M, given by $z_j = \mu_{z_0,\theta_j}(s_0)$, where $\mu_{z_0,\theta_j}(s)$ are geodesics of $(\partial M, g_{\partial M})$ and $\theta_1,\dots,\theta_{n-1}$ are orthonormal vectors of $T_{z_0}(\partial M)$ with respect to metric $g_{\partial M}$ and $s_0 > 0$ is sufficiently small, so that the distance functions $y \mapsto d_g(z_i,y)$, $i = 0,1,2,\dots,n-1$ form local coordinates $y \mapsto \left(d_g(z_i,y)\right)_{i=0}^{n-1}$ on M in some neighborhood of the point $x^{(0)}$ (we omit here the proof which can be found in [42, Lemma 2.14]).

Let now $W \subset R(M)$ be a neighborhood of $r^{(0)}$ and let $\tilde{r} \in W$. Moreover, let $V = R^{-1}(W) \subset M$ and $\tilde{x} = R^{-1}(\tilde{r}) \in V$. Let us next consider arbitrary points $z_1,\dots,z_{n-1}$ on ∂M. Our aim is to verify whether the functions $x \mapsto X^i(x) = d_g(x,z_i)$, $i = 0,1,\dots,n-1$ form smooth coordinates in V. As $M\backslash\omega$ is dense on M and we have found topological structure of $R(M)$ and constructed the metric g_R on $R(M\backslash\omega)$, we can choose $r^{(j)} \in R(M\backslash\omega)$ such that $\lim_{j\to\infty} r^{(j)} = \tilde{r}$ in $R(M)$. Let $x^{(j)} \in M\backslash\omega$ be the points for which $r^{(j)} = R\left(x^{(j)}\right)$. Now the function $x \mapsto \left(X^i(x)\right)_{i=0}^{n-1}$ defines smooth coordinates near $\tilde{x}$ if and only if for functions $Z^i(r) = K_{z_i}(r)$, we have

$$\begin{aligned}\lim_{j\to\infty} \det\left(\left(g_R\left(dZ^i(r), dZ^l(r)\right)\right)_{i,l=0}^{n-1}\right)|_{r=r^{(j)}} \\ = \lim_{j\to\infty} \det\left(\left(g\left(dX^i(x), dX^l(x)\right)\right)_{i,l=0}^{n-1}\right)|_{x=x^{(j)}} \neq 0.\end{aligned} \tag{23}$$

Thus, for all $\tilde{r} \in W$ we can verify for any points $z_1,\dots,z_{n-1} \in \partial M$ whether the condition (23) is valid or not and this condition is valid for all $\tilde{r} \in W$ if and only if the functions $x \mapsto X^i(x) = d_g(x,z_i)$, $i = 0,1,\dots,n-1$ form smooth coordinates in V. Moreover, by the above reasoning, we know that any $r^{(0)} \in R(\omega)$ has some neighborhood W and some points $z_1,\dots,z_{n-1} \in \partial M$ for which the condition (23) is valid for all $\tilde{r} \in W$. By choosing such points, we find also near $r^{(0)} \in (\omega)$ smooth coordinates $r \mapsto \left(Z^i(r)\right)_{i=0}^{n-1}$ which make the map $R : M \to R(M)$ a diffeomorphism near $x^{(0)}$.

Summarizing, we have constructed differentiable structure (i.e., local coordinates) on the whole set $R(M)$, and this differentiable structure makes the map $R : M \to R(M)$ a diffeomorphism. Moreover, since the metric $g_R = R_* g$ is a smooth tensor, and we have found it in a dense subset $R(M \setminus \omega)$ of $R(M)$, we can continue it in the local coordinates. This gives us the metric g_R on the whole $R(M)$, which makes the map $R : M \to R(M)$ an isometric diffeomorphism. ■

In the above proof, the reconstruction of the metric tensor in the boundary normal coordinates can be considered as finding the image of the metric in the travel time coordinates.

Let us next consider the case when we have an unknown isotropic wave speed $c(x)$ in a bounded domain $\Omega \subset \mathbb{R}^n$. We will assume that we are given the set Ω and an abstract Riemannian manifold (M, g), which is isometric to Ω endowed with its travel time metric corresponding to the wave speed $c(x)$. Also, we assume that we are given a map $\psi : \partial\Omega \to \partial M$, which gives the correspondence between the boundary points of Ω and M. Next we show that it is then possible to find an embedding from the manifold M to Ω which gives us the wave speed $c(x)$ at each point $x \in \Omega$. This construction is presented in detail, e.g., in [42].

For this end, we need first to reconstruct a function σ on M which corresponds to the function $c(x)^2$ on Ω. This is done on the following lemma.

Lemma 2. *Assume we are given a Riemannian manifold (M, g) such that there exists an open set $\Omega \subset \mathbb{R}^n$ and an isometry $\Psi : \big(\Omega, \big(\sigma(x)\big)^{-1}\delta_{ij}\big) \to (M, g)$ and a function α on M such that $\alpha\big(\Psi(x)\big) = \sigma(x)$. Then knowing the Riemannian manifold (M, g), the restriction $\psi = \Psi|_{\partial\Omega} : \partial\Omega \to \partial M$, and the boundary value $\sigma|_{\partial\Omega}$, we can determine the function α.*

Proof. First, observe that we are given the boundary value $\alpha|_{\partial M}$ of $\alpha\big(\Psi(x)\big) = \sigma(x)$. By assumption, the metric g on M is conformally Euclidean, that is, the metric tensor, in some coordinates, has the form $g_{jk}(x) = \sigma(x)^{-1}\delta_{jk}$, where $\sigma(x) > 0$. Hence, the function $\beta = \frac{1}{2}\ln(\alpha)$, when $m = 2$, and $\beta = \alpha^{(n-2)/4}$, when $n \geq 3$, satisfies the so-called scalar curvature equation

$$\Delta_g \beta - k_g = 0 \quad (n = 2), \tag{24}$$

$$\frac{4(n-1)}{n-2}\Delta_g \beta - k_g \beta = 0 \quad (n \geq 3), \tag{25}$$

where k_g is the scalar curvature of (M, g),

$$k_g(x) = \sum_{k,j,l=1}^{n} g^{jl}(x) R^k_{jkl}(x)$$

where R^i_{jkl} is the curvature tensor given in terms of the Christoffel symbols as

$$R^i_{jkl}(x) = \frac{\partial}{\partial x^k}\Gamma^i_{lj}(x) - \frac{\partial}{\partial x^l}\Gamma^i_{kj}(x) + \sum_{r=1}^{n}\left(\Gamma^r_{lj}(x)\Gamma^i_{kr}(x) - \Gamma^r_{kj}(x)\Gamma^i_{lr}(x)\right).$$

The idea of these equations is that if β satisfies, for example, Eq. (25) in the case $m \geq 3$, then the metric $\beta^{4/(n-2)}g$ has zero scalar curvature. Together with boundary data (10) being given, we obtain Dirichlet boundary value problem for β in M.

Clearly, Dirichlet problem for Eq. (24) has a unique solution that gives α when $n = 2$. In the case $n \geq 3$, to show that this boundary value problem has a unique solution, it is necessary to check that 0 is not an eigenvalue of the operator $\frac{4(n-1)}{n-2}\Delta_g - k_g$ with Dirichlet boundary condition. Now, the function $\beta = \alpha^{(n-2)/4}$ is a positive solution of the Dirichlet problem for Eq. (25) with boundary condition $\beta|_{\partial M} = \alpha^{(n-2)/4}\big|_{\partial M}$. Assume that there is another possible solution of this problem,

$$\tilde{\beta} = v\beta, \quad v > 0, \quad v|_{\partial M} = 1. \tag{26}$$

Then both $\left(M, \beta^{4/(n-2)}g\right)$ and $\left(M, \tilde{\beta}^{4/(n-2)}g\right)$ have zero scalar curvatures. Denoting $g_1 = \beta^{4/(n-2)}g$, $g_2 = \tilde{\beta}^{4/(n-2)}g$, we obtain that v should satisfy the scalar curvature equation

$$\frac{4(n-1)}{n-2}\Delta_{g_1}v - k_{g_1}v = 0.$$

Here, we have $k_{g_1} = 0$ as g_1 has vanishing scalar curvature. Together with boundary condition (26), this equation implies that $v \equiv 1$, that is, $\beta = \tilde{\beta}$. This immediately yields that 0 is not the eigenvalue of the Dirichlet operator (25) because, otherwise, we could obtain a positive solution $\tilde{\beta} = \beta + c_0\psi_0$, where ψ_0 is the Dirichlet eigenfunction, corresponding to zero eigenvalue, and $|c_0|$ is sufficiently small. Thus, β, and henceforth α, can be uniquely determined by solving Dirichlet boundary value problems for (24) and (25). ■

Our next goal is to embed the abstract manifold (M, g) with conformally Euclidean metric into Ω with metric $\left(\sigma(x)\right)^{-1}\delta_{ij}$. To achieve this goal, we use the a priori knowledge that such embedding exists and the fact that we have already constructed α corresponding to $\sigma(x)$ on M.

Lemma 3. *Let (M, g) be a compact Riemannian manifold, $\alpha(x)$ a positive smooth function on M, and $\psi : \partial\Omega \to \partial M$ a diffeomorphism. Assume also that there is a diffeomorphism $\Psi : \overline{\Omega} \to M$ such that*

$$\Psi|_{\partial\Omega} = \psi, \quad \Psi^* g = \left(\alpha\left(\Psi(x)\right)\right)^{-1}\delta_{ij}.$$

Then, if Ω, (M, g), α, and ψ are known, it is possible to construct the diffeomorphism Ψ by solving ordinary differential equations.

Proof. Let $\zeta = (z, \tau)$ be the boundary normal coordinates on $M \setminus \omega$. Our goal is to construct the coordinate representation for $\Psi^{-1} = X$,

$$X : M \setminus \omega \to \Omega,$$
$$X(z, \tau) = \big(x^1(z, \tau), \dots, x^n(z, \tau)\big).$$

Denote by $h_{ij}(x) = \alpha\big(\Psi(x)\big)^{-1}\delta_{ij}$ the metric tensor in Ω. Let $\Gamma_{i,jk} = \sum_p g_{ip}\Gamma^p_{jk}$ be the Christoffel symbols of (Ω, h_{ij}) in the Euclidean coordinates and let $\tilde{\Gamma}_{\sigma,\mu\nu}$ be Christoffel symbols of (M, g), in ζ-coordinates. Next, we consider functions h_{ij}, $\Gamma_{k,ij}$, etc., as functions on $M \setminus \omega$ in (z, τ)-coordinates evaluated at the point $x = x(z, \tau)$, for example, $\Gamma_{k,ij}(z, \tau) = \Gamma_{k,ij}\big(x(z, \tau)\big)$. Then, since Ψ is an isometry, the transformation rule of Christoffel symbols with respect to the change of coordinates implies

$$\tilde{\Gamma}_{\sigma,\mu\nu} = \sum_{i,j,k=1}^{n} \Gamma_{k,ij} \frac{\partial x^i}{\partial \zeta^\mu} \frac{\partial x^j}{\partial \zeta^\nu} \frac{\partial x^k}{\partial \zeta^\sigma} + \sum_{i,j=1}^{n} h_{ij} \frac{\partial x^i}{\partial \zeta^\sigma} \frac{\partial^2 x^j}{\partial \zeta^\mu \partial \zeta^\nu}, \tag{27}$$

where

$$h_{ij}(z, \tau) = \frac{1}{\alpha\big(\Psi(z, \tau)\big)}\delta_{ij}. \tag{28}$$

Using Eqs. (27) and (28), we can write $\frac{\partial^2 x^j}{\partial \zeta^\mu \partial \zeta^\nu}$ in the form

$$\begin{aligned} \frac{\partial^2 x^j}{\partial \zeta^\mu \partial \zeta^\nu}(\zeta) \;=\; & \sum_{p,\sigma,\mu,\nu=1}^{n} \alpha(\zeta)\delta^{jp}\Bigg(\tilde{\Gamma}_{\sigma,\mu\nu}\frac{\partial \zeta^\sigma}{\partial x^p} - \sum_{n=1}^{n}\frac{1}{2}\frac{\partial \alpha^{-1}}{\partial \zeta^\sigma} \\ & \times \left[\frac{\partial \zeta^\sigma}{\partial x^n}\delta_{pi} + \frac{\partial \zeta^\sigma}{\partial x^i}\delta_{pn} - \frac{\partial \zeta^\sigma}{\partial x^p}\delta_{ni}\right]\frac{\partial x^i}{\partial \zeta^\mu}\frac{\partial x^n}{\partial \zeta^\nu}\Bigg). \end{aligned} \tag{29}$$

As α and $\tilde{\Gamma}_{\sigma,\mu\nu}$ are known as a function of ζ, the right-hand side of (29) can be written in the form

$$\frac{\partial^2 x^j}{\partial \zeta^\mu \partial \zeta^\nu} = F^j_{\mu,\nu}\left(\zeta, \frac{\partial x}{\partial \zeta}\right), \tag{30}$$

where $F^j_{\mu,\nu}$ are known functions. Choose $\nu = m$, so that

$$\frac{\partial^2 x^j}{\partial \zeta^\mu \partial \zeta^n} = \frac{d}{d\tau}\left(\frac{\partial x^j}{\partial \zeta^\mu}\right).$$

Then, Eq. (30) becomes a system of ordinary differential equations along normal geodesics for the matrix $\left(\frac{\partial x^j}{\partial \zeta^\mu}(\tau)\right)_{j,\mu=1}^{n}$. Moreover, since diffeomorphism $\Psi : \partial\Omega \to \partial M$ is given, the boundary derivatives $\frac{\partial x^j}{\partial \zeta^\mu}$, $\mu = 1, \ldots, n-1$, are known for $\zeta^n = \tau = 0$. By relation (28),

$$\frac{\partial x^j}{\partial \zeta^n} = \frac{\partial x^j}{\partial \tau} = \alpha^{-1} \frac{\partial x^j}{\partial \mathbf{n}} = -\alpha^{-1} \mathbf{n}^j$$

for $\zeta^n = \tau = 0$ where $\mathbf{n} = (\mathbf{n}^1, \ldots, \mathbf{n}^n)$ is the Euclidean unit exterior normal vector. Thus, $\frac{\partial x^j}{\partial \tau}(z, 0)$ are also known. Solving a system of ordinary differential equations (30) with these initial conditions at $\tau = 0$, we can construct $\frac{\partial x^j}{\partial \zeta^\mu}(z, \tau)$ everywhere on $M \setminus \omega$. In particular, taking $\mu = n$, we find $\frac{dx^j}{d\tau}(z, \tau)$. Using again the fact that $\big(x^1(z, 0), \ldots, x^n(z, 0)\big) = \psi(z)$ are known, we obtain the functions $x^j(z, \tau)$, z fixed, $0 \le \tau \le \tau_{\partial M}(z)$, that is, reconstruct all normal geodesics on Ω with respect to metric h_{ij}. Clearly, this gives us the embedding of (M, g) onto (Ω, h_{ij}). ■

Combining the above results, we get the following result for the isotropic wave equation.

Theorem 2. *Let $\Omega \subset \mathbb{R}^n$ to be a bounded, open set with smooth boundary and $c(x) \in C^\infty(\overline{\Omega})$ be a strictly positive function. Assume that we know Ω, $c|_{\partial\Omega}$, and the nonstationary Robin-to-Neumann map $\Lambda_{\partial\Omega}$. Then it is possible to determine the function $c(x)$.*

We note that in Theorem 2, the boundary value $c|_{\partial\Omega}$ of the wave speed $c(x)$ can be determined using the finite velocity of wave propagation (9) and the knowledge of Ω and $\Lambda_{\partial\Omega}$, but we will not consider this fact in this chapter.

From Boundary Data to Inner Products of Waves

Let $u^f(x, t)$ denote the solutions of the hyperbolic equation (3), Λ^{2T} be the finite time Robin-to-Dirichlet map for Eq. (3) and let dS_g denote the Riemannian volume form on the manifold $(\partial M, g_{\partial M})$. We start with the Blagovestchenskii identity.

Lemma 4. *Let $f, h \in C_0^\infty(\partial M \times \mathbb{R}_+)$. Then*

$$\int_M u^f(x, T) u^h(x, T)\, dV_\mu(x) = \tag{31}$$

$$= \frac{1}{2} \int_L \int_{\partial M} \big(f(x,t)(\Lambda^{2T} h)(x,s) - (\Lambda^{2T} f)(x,t) h(x,s) \big)\, dS_g(x) dt\, ds,$$

where

$$L = \{(s,t) : \ 0 \le t + s \le 2T, \ t < s, \ t, s > 0\}.$$

Proof. Let

$$w(t,s) = \int_M u^f(x,t) u^h(x,s) \, dV_\mu(x).$$

Then, by integration by parts, we see that

$$\begin{aligned}
\left(\partial_t^2 - \partial_s^2\right) w(t,s) &= \int_M \left[\partial_t^2 u^f(x,t) u^h(x,s) - u^f(x,t) \partial_s^2 u^h(x,s)\right] dV_\mu(x) = \\
&= -\int_M \left[A u^f(x,t) u^h(x,s) - u^f(x,t) A u^h(x,s)\right] dV_\mu(x) = \\
&= -\int_{\partial M} \left[B_{\nu,\eta} u^f(t) u^h(s) - u^f(t) B_{\nu,\eta} u^h(s)\right] dS_g(x) = \\
&= \int_{\partial M} \left[\Lambda^{2T} u^f(x,t) u^h(x,s) - u^f(x,t) \Lambda^{2T} u^h(x,s)\right] dS_g(x).
\end{aligned}$$

Moreover,

$$\begin{aligned}
w|_{t=0} &= w|_{s=0} = 0, \\
\partial_t w|_{t=0} &= \partial_s w|_{s=0} = 0.
\end{aligned}$$

Thus, w is the solution of the initial boundary value problem for the one-dimensional wave equation in the domain $(t,s) \in [0,2T] \times [0,2T]$ with known source and zero initial and boundary data (10). Solving this problem, we determine $w(t,s)$ in the domain where $t + s \le 2T$ and $t < s$ (see Fig. 1). In particular, $w(T,T)$ gives the assertion. ■

The other result is based on the following fundamental theorem by D. Tataru [77, 79].

Theorem 3. *Let $u(x,t)$ solve the wave equation $u_{tt} + Au = 0$ in $M \times \mathbb{R}$ and $u|_{\Gamma \times (0,2T_1)} = \partial_\nu u|_{\Gamma \times (0,2T_1)} = 0$, where $\emptyset \neq \Gamma \subset \partial M$ is open. Then*

$$u = 0 \text{ in } K_{\Gamma,T_1}, \tag{32}$$

where

$$K_{\Gamma,T_1} = \left\{(x,t) \in M \times \mathbb{R} : d_g(x,\Gamma) < T_1 - |t - T_1|\right\}$$

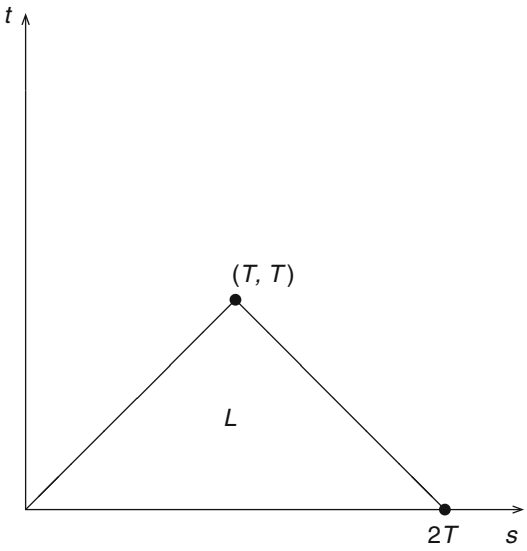


Fig. 1 Domain of integration in the Blagovestchenskii identity

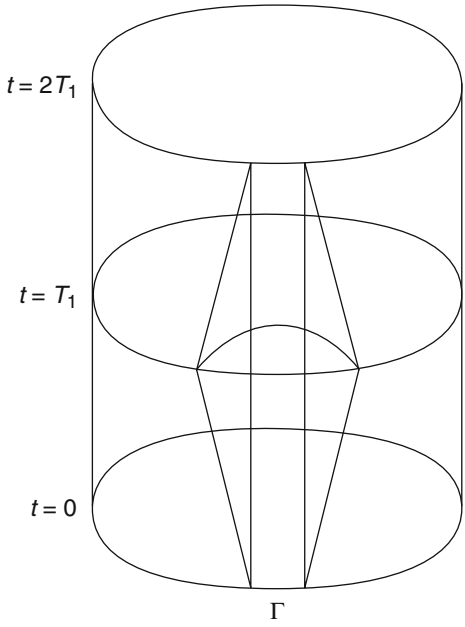


Fig. 2 Double cone of influence

is the double cone of influence (see Fig. 2).

(The proof of this theorem, in full generality, is in [77]. A simplified proof for the considered case is in [42].)

The observability Theorem 3 gives rise to the following approximate controllability:

Corollary 1. *For any open* $\Gamma \subset \partial M$ *and* $T_1 > 0$,

$$\mathrm{cl}_{L^2(M)} \left\{ u^f(\cdot, T_1) : f \in C_0^\infty\big(\Gamma \times (0, T_1)\big) \right\} = L^2\big(M(\Gamma, T_1)\big).$$

Here,

$$M(\Gamma, T_1) = \{x \in M : d_g(x, \Gamma) < T_1\} = K_{\Gamma, T_1} \cap \{t = T_1\}$$

is the domain of influence of Γ at time T_1 and $L^2\big(M(\Gamma, T_1)\big) = \{a \in L^2(M) : \mathrm{supp}(a) \subset M(\Gamma, T_1)\}$.

Proof. Let us assume that $a \in L^2\big(M(\Gamma, T_1)\big)$ is orthogonal to all $u^f(\cdot, T_1)$, $f \in C_0^\infty(\Gamma \times (0, T_1))$. Denote by v the solution of the wave equation

$$\begin{aligned} &(\partial_t^2 + A)v = 0; \quad v|_{t=T_1} = 0, \quad \text{in } M \times \mathbb{R}, \\ &\partial_t v|_{t=T_1} = a; \quad B_{v,\eta} v|_{\partial M \times \mathbb{R}} = 0. \end{aligned}$$

Using integration by parts, we obtain for all $f \in C_0^\infty\big(\Gamma \times (0, T_1)\big)$

$$\int_0^{T_1} \int_{\partial M} f(x, s) v(x, s)\, dS_g(x)\, ds = \int_M a(x) u^f(x, T_1) dV_\mu(x) = 0,$$

due to the orthogonality of a and the solutions $u^f(t)$. Thus, $v|_{\Gamma \times (0, T_1)} = 0$. Moreover, as v is odd with respect to $t = T_1$, that is, $v(x, T_1 + s) = -v(x, T_1 - s)$, we see that $v|_{\Gamma \times (T_1, 2T_1)} = 0$. As u satisfies the wave equation, standard energy estimates yield that $u \in C\big(\mathbb{R}; H^1(M)\big)$, and hence $u|_{\partial M \times \mathbb{R}} \in C\big(\mathbb{R}; H^{1/2}(\partial M)\big)$. Combining the above, we see that $v|_{\Gamma \times (0, 2T_1)} = 0$, and as $B_{v,\eta} v|_{\Gamma \times (0, 2T_1)} = 0$, we see using Theorem 3 that $a = 0$. ∎

Recall that we denote $u^f(t) = u^f(\cdot, t)$.

Lemma 5. *Let* $T > 0$ *and* $\Gamma_j \subset \partial M$, $j = 1, \ldots, J$, *be nonempty, relatively compact open sets,* $0 \le T_j^- < T_j^+ \le T$. *Assume we are given* $(\partial M, g_{\partial M})$ *and the response operator* Λ^{2T}. *This data determines the inner product*

$$J_N^T(f_1, f_2) = \int_N u^{f_1}(x, t) u^{f_2}(x, t)\, dV_\mu(x)$$

for given $t > 0$ *and* $f_1, f_2 \in C_0^\infty(\partial M \times \mathbb{R}_+)$, *where*

$$N = \bigcap_{j=1}^J \Big(M\left(\Gamma_j, T_j^+\right) \setminus M\left(\Gamma_j, T_j^-\right) \Big) \subset M.$$

Proof. Let us start with the case when $f_1 = f_2 = f$ and $T_j^- = 0$ for all $j = 1, 2, \ldots, J$.

Let $B = \bigcup_{j=1}^{J}(\Gamma_j \times [T - T_j, T])$. For all $h \in C_0^\infty(B)$, it holds by (9) that $\text{supp}\big(u^h(\cdot, T)\big) \subset N$, and thus

$$\begin{aligned}&\|u^f(T) - u^h(T)\|^2_{L^2(M,dV_\mu)}\\ &= \int_N \big(u^f(x,T) - u^h(x,T)\big)^2\, dV_\mu(x) + \int_{M\setminus N} \big(u^f(x,T)\big)^2\, dV_\mu(x).\end{aligned}$$

Let $\chi_N(x)$ be the characteristic function of the set N. By Corollary 1, there is $h \in C_0^\infty(B)$ such that the norm $\|\chi_N u^f(T) - u^h(T)\|_{L^2(M,dV_\mu)}$ is arbitrarily small. This shows that $J_N^T(f_1, f_2)$ can be found by

$$J_N^T(f, f) = \|u^f(T)\|^2_{L^2(M,dV_\mu)} - \inf_{h\in C_0^\infty(B)} F(h), \tag{33}$$

where

$$F(h) = \|u^f(T) - u^h(T)\|^2_{L^2(M,dV_\mu)}.$$

As $F(h)$ can be computed with the given data (10) by Lemma 4, it follows that we can determine $J_N^T(f, f)$ for any $f \in C_0^\infty(\partial M \times \mathbb{R}_+)$. Now, since

$$J_N^T(f_1, f_2) = \frac{1}{4}\left(J_N^T(f_1 + f_2, f_1 + f_2) - J_N^T(f_1 - f_2, f_1 - f_2)\right),$$

the claim follows in the case when $T_j^- = 0$ for all $j = 1, 2, \ldots, J$.

Let us consider the general case when T_j^- may be nonzero. We observe that we can write the characteristic function $\chi_N(x)$ of the set $N = \bigcap_{j=1}^{J}\left(M\left(\Gamma_j, T_j^+\right)\Big\backslash M\left(\Gamma_j, T_j^-\right)\right)$ as

$$\chi_N(x) = \sum_{k=1}^{K_1} c_k \chi_{N_k}(x) - \sum_{k=K_1+1}^{K_2} c_k \chi_{N_k}(x),$$

where $c_k \in \mathbb{R}$ are constants which can be determined by solving a simple linear system of equations and the sets N_k are of the form

$$N_k = \bigcup_{j\in I_k} M(\Gamma_j, t_j),$$

where $I_k \subset \{1,2,\ldots,J\}$ and $t_j \in \left\{T_j^+ : j = 1,2,\ldots,J\right\} \cup \left\{T_j^- : j = 1,2,\ldots,J\right\}$. Thus,

$$J_N^T(f_1,f_2) = \sum_{k=1}^{K_1} c_k J_{N_k}^T(f_1,f_2) - \sum_{k=K_1+1}^{K_2} c_k J_{N_k}^T(f_1,f_2),$$

where all the terms $J_{N_k}^T(f_1,f_2)$ can be computed using the boundary data (10). ■

From Inner Products of Waves to Boundary Distance Functions

Let us consider open sets $\Gamma_j \subset \partial M$, $j = 1,2,\ldots,J$ and numbers $T_j^+ > T_j^- \geq 0$. For a collection $\left\{\left(\Gamma_j, T_j^+, T_j^-\right) : j = 1,\ldots,J\right\}$, we define the number

$$P\left(\left\{\left(\Gamma_j, T_j^+, T_j^-\right) : j = 1,\ldots,J\right\}\right) = \sup_f J_N^T(f,f),$$

where $T = \left(\max T_j^+\right) + 1$,

$$N = \bigcap_{j=1}^{J} \left(M\left(\Gamma_j, T_j^+\right) \setminus M\left(\Gamma_j, T_j^-\right)\right)$$

and the supremum is taken over functions $f \in C_0^\infty\left(\partial M \times (0,T)\right)$ satisfying $\|u^f(T)\|_{L^2(M)} \leq 1$. When $\Gamma_j^q \subset \partial M$, $j = 1,2,\ldots,J$, are open sets, so that $\Gamma_j^q \to \{z_j\}$ as $q \to \infty$, that is, $\{z_j\} \subset \Gamma_j^q \subset \Gamma_j^{q-1}$ for all q and $\bigcap_{q=1}^\infty \overline{\Gamma}_j^q = \{z_j\}$, we denote

$$P\left(\left\{\left(z_j, T_j^+, T_j^-\right) : j = 1,\ldots,J\right\}\right) = \lim_{q\to\infty} P\left(\left\{\left(\Gamma_j^q, T_j^+, T_j^-\right) : j = 1,\ldots,J\right\}\right).$$

Theorem 4. *Let $\{z_n\}_{n=1}^\infty$ be a dense set on ∂M and $r(\cdot) \in C(\partial M)$ be an arbitrary continuous function. Then $r \in R(M)$ if and only if for all $N > 0$ it holds that*

$$P\left(\left\{\left(z_j, r(z_n) + \frac{1}{N}, r(z_n) - \frac{1}{N}\right) : j = 1,\ldots,N\right\}\right) > 0. \tag{34}$$

Moreover, condition (34) can be verified using the boundary data (10). Hence, the boundary data determine uniquely the boundary distance representation $R(M)$ of (M,g) and therefore determine the isometry type of (M,g).

Proof. "If"–part. Let $x \in M$ and denote for simplicity $r(\cdot) = r_x(\cdot)$. Consider a ball $B_{1/N}(x) \subset M$ of radius $1/N$ and center x in (M,g). Then, for $z \in \partial M$

$$B_{1/N}(x) \subset M\left(z, r(z)+\frac{1}{N}\right)\setminus M\left(z, r(z)-\frac{1}{N}\right).$$

By Corollary 1, for any $T > r(z)$, there is $f \in C_0^\infty\big(\partial M \times (0,T)\big)$ such that the function $u^f(\cdot, T)$ does not vanish a.e. in $B_{1/N}(x)$. Thus, for any $N \in \mathbb{Z}_+$ and $T = \max\{r(z_n) : n = 1, 2, \ldots, N\}$, we have

$$P\left(\left\{\left(z_j, r(z_n)+\frac{1}{N}, r(z_n)-\frac{1}{N}\right) : j = 1, \ldots, N\right\}\right)$$
$$\geq \int_{B_{1/N}(x)} |u^f(x,T)|^2\, dV_\mu(x) > 0$$

"Only if"–part. Let (34) be valid. Then for all $N > 0$, there are points

$$x_N \in A_N = \bigcap_{n=1}^{N}\left(M\left(z_n, r(z_n)+\frac{1}{N}\right)\setminus M\left(z_n, r(z_n)-\frac{1}{N}\right)\right) \tag{35}$$

as the set A_N has to have a nonzero measure. By choosing a suitable subsequence of x_N (denoted also by x_N), there exists a limit $x = \lim_{N\to\infty} x_N$.

Let $j \in \mathbb{Z}_+$. It follows from (35) that

$$r(z_j) - \frac{1}{N} \leq d_g(x_N, z_j) \leq r(z_j) + \frac{1}{N} \quad \text{for all } N \geq j.$$

As the distance function d_g on M is continuous, we see by taking limit $N \to \infty$ that

$$d_g(x, z_j) = r(z_j), \quad j = 1, 2, \ldots.$$

Since $\{z_j\}_{j=1}^\infty$ are dense in ∂M, we see that $r(z) = d_g(x,z)$ for all $z \in \partial M$, that is, $r = r_x$. ∎

Note that this proof provides an algorithm for construction of an isometric copy of (M, g) when the boundary data (10) are given.

Alternative Reconstruction of Metric via Gaussian Beams

Next we consider an alternative construction of the boundary distance representation $R(M)$, developed in [6,41,42]. In the previous considerations, we used in Lemma 5 the sets of type $N = \bigcap_{j=1}^J \left(M\left(\Gamma_j, T_j^+\right)\Big\setminus M\left(\Gamma_j, T_j^-\right)\right) \subset M$ and studied the norms $\|\chi_N u^f(\cdot, T)\|_{L^2(M)}$. In the alternative construction considered below, we need to consider only the sets N of the form $N = M(\Gamma_0, T_0)$. For this end, we consider solutions $u^f(x,t)$ with special sources f which produce wave packets, called the Gaussian beams [3,63]. For simplicity, we consider just the case when

$$A = -\Delta_g + q,$$

and give a very short exposition on the construction of the Gaussian beam solutions. Details can be found in, for example, Ref. [42], Chapter 2.4, where the properties of Gaussian beams are discussed in detail. In this section, we consider complex valued solutions $u^f(x,t)$.

Gaussian beams, called also "quasiphotons," are a special class of solutions of the wave equation depending on a parameter $\varepsilon > 0$ which propagate in a neighborhood of a geodesic $\gamma = \gamma_{y,\xi}([0,L])$, $g(\xi,\xi) = 1$. Below, we consider first the construction in the case when γ is in the interior of M.

To construct Gaussian beams, we start by considering an asymptotic sum, called formal Gaussian beam,

$$U_\varepsilon(x,t) = M_\varepsilon \exp\{-(i\varepsilon)^{-1}\theta(x,t)\} \sum_{k=0}^{N} u_k(x,t)(i\varepsilon)^k, \tag{36}$$

where $x \in M$, $t \in [t_-,t_+]$, and $M_\varepsilon = (\pi\varepsilon)^{-n/4}$ is the normalization constant. The function $\theta(x,t)$ is called the phase function and $u_k(x,t)$, $k = 0,1,\ldots,N$ are the amplitude functions. A phase function $\theta(x,t)$ is associated with a geodesic $t \mapsto \gamma(t) \in M$ if

$$\operatorname{Im}\theta\big(\gamma(t),t\big) = 0, \tag{37}$$

$$\operatorname{Im}\theta(x,t) \geq C_0 d_g\big(x,\gamma(t)\big)^2, \tag{38}$$

for $t \in [t_-,t_+]$. These conditions guarantee that for any t, the absolute value of $U_\varepsilon(x,t)$ looks like a Gaussian function in the x variable which is centered at $\gamma(t)$. Thus, the formal Gaussian beam can be considered to move in time along the geodesic $\gamma(t)$. The phase function can be constructed, so that it satisfies the eikonal equation

$$\left(\frac{\partial}{\partial t}\theta(x,t)\right)^2 - g^{jl}(x)\frac{\partial}{\partial x^j}\theta(x,t)\,\frac{\partial}{\partial x^l}\theta(x,t) \asymp 0, \tag{39}$$

where $\asymp$ means the coincidence of the Taylor coefficients of both sides considered as functions of x at the points $\gamma(t), t \in [t_-,t_+]$, that is,

$$v(x,t) \asymp 0 \quad \text{if } \partial_x^\alpha v(x,t)|_{x=\gamma(t)} = 0 \text{ for all } \alpha \in \mathbb{N}^n \text{ and } t \in [t_-,t_+].$$

The amplitude functions u_k, $k = 0,\ldots,N$ can be constructed as solutions of the transport equations

$$\mathcal{L}_\theta u_k \asymp \big(\partial_t^2 - \Delta_g + q\big)\,u_{k-1}, \quad \text{with } u_{-1} = 0. \tag{40}$$

Here $\mathcal{L}_\theta$ is the transport operator

$$\mathcal{L}_\theta u = 2\partial_t \theta \partial_t u - 2\langle \nabla\theta, \nabla u\rangle_g + \left(\partial_t^2 - \Delta_g\right)\theta \cdot u, \tag{41}$$

where $\nabla u(x,t) = \sum_j g^{jk}(x)\frac{\partial u}{\partial x^k}(x,t)\frac{\partial}{\partial x^k}$ is the gradient on (M,g) and $\langle V, W\rangle_g = \sum_{j=1}^n g^{jk}(x)V_j(x)W_k(x)$. The following existence result is proven, for example, in [3,42,63].

Theorem 5. *Let $y \in M^{\text{int}}$, $\xi \in T_xM$ be a unit vector and $\gamma = \gamma_{y,\xi}(t)$, $t \in [t_-, t_+] \subset \mathbb{R}$ be a geodesic lying in M^{int} when $t \in (t_-, t_+)$.*

Then there are functions $\theta(x,t)$ and $u_k(x,t)$ satisfying (38)–(40) and a solution $u_\varepsilon(x,t)$ of equation

$$\left(\partial_t^2 - \Delta_g + q\right)u_\varepsilon(x,t) = 0, \quad (x,t) \in M \times [t_-, t_+], \tag{42}$$

such that

$$|u_\varepsilon(x,t) - \phi(x,t)U_\varepsilon(x,t)| \le C_N \varepsilon^{\tilde{N}(N)}, \tag{43}$$

where $\tilde{N}(N) \to \infty$ when $N \to \infty$. Here $\phi \in C_0^\infty(M \times \mathbb{R})$ is a smooth cut-off function satisfying $\phi = 1$ near the trajectory $\{(\gamma(t), t) : t \in [t_-, t_+]\} \subset M \times \mathbb{R}$.

In the other words, for an arbitrary geodesic in the interior of M, there is a Gaussian beam that propagates along this geodesic.

Next we consider a class of boundary sources in (3) which generate Gaussian beams. Let $z_0 \in \partial M$, $t_0 > 0$, and let $x \mapsto z(x) = \left(z^1(x), \dots, z^{n-1}(x)\right)$ be a local system of coordinates on $W \subset \partial M$ near z_0. For simplicity, we denote these coordinates as $z = (z^1, \dots, z^{n-1})$ and make computations without reference to the point x. Consider a class of functions $f_\varepsilon = f_{\varepsilon,z_0,t_0}(z,t)$ on the boundary cylinder $\partial M \times \mathbb{R}$, where

$$f_\varepsilon(z,t) = B_{v,\eta}\left((\pi\varepsilon)^{-n/4}\phi(z,t)\exp\{i\varepsilon^{-1}\Theta(z,t)\}V(z,t)\right). \tag{44}$$

Here $\phi \in C_0^\infty(\partial M \times \mathbb{R})$ is one near (z_0, t_0) and

$$\Theta(z,t) = -(t-t_0) + \frac{1}{2}\langle H_0(z-z_0), (z-z_0)\rangle + \frac{i}{2}(t-t_0)^2, \tag{45}$$

where $\langle\cdot,\cdot\rangle$ is the complexified Euclidean inner product, $\langle a, b\rangle = \sum a_j b_j$, and $H_0 \in \mathbb{C}^{n\times n}$ is a symmetric matrix with a positive definite imaginary part, that is, $(H_0)_{jk} = (H_0)_{kj}$ and $\operatorname{Im} H_0 > 0$, where $(\operatorname{Im} H_0)_{jk} = \operatorname{Im}(H_0)_{jk}$. Finally, $V(z,t)$ is a smooth function supported in $W \times \mathbb{R}_+$, having nonzero value at (z_0, t_0). The solution $u^{f_\varepsilon}(x,t)$ of the wave equation

$$\begin{aligned} &\partial_t^2 u - \Delta_g u + qu = 0, \quad \text{in } M \times \mathbb{R}_+, \\ &u|_{t=0} = \partial_t u|_{t=0} = 0, \\ &B_{\nu,\eta} u\big|_{\partial M \times \mathbb{R}_+} = f_\varepsilon(z,t) \end{aligned} \tag{46}$$

is a Gaussian beam propagating along the normal geodesic $\gamma_{z_0,\nu}$. Let $S(z_0) \in (0,\infty]$ be the smallest values $s > 0$, so that $\gamma_{z_0,\nu}(s) \in \partial M$, that is, the first time when the geodesic $\gamma_{z_0,\nu}$ hits to ∂M, or $S(z_0) = \infty$ if no such value $s > 0$ exists. Then the following result in valid (see, e.g., [42]).

Lemma 6. *For any function $V \in C_0^\infty(W \times \mathbb{R}_+)$ being one near (z_0, t_0), $t_0 > 0$, and $0 < t_1 < S(z_0)$ and $N \in \mathbb{Z}_+$, there are C_N so that the solution $u^{f_\varepsilon}(x,t)$ of problem (46) satisfies estimates*

$$|u^{f_\varepsilon}(x,t) - \phi(x,t)U_\varepsilon(x,t)| \le C_N \varepsilon^{\tilde{N}(N)}, \quad 0 \le t < t_0 + t_1 \tag{47}$$

where $U_\varepsilon(x,t)$ is of the form (36), for all $0 < \varepsilon < 1$, where $\tilde{N}(N) \to \infty$ when $N \to \infty$ and $\phi \in C_0^\infty(M \times \mathbb{R})$ is ϕ one near the trajectory $\{(\gamma_{z_0,\nu}(t), t + t_0) : t \in [0, t_1]\} \subset M \times \mathbb{R}$.

Let us denote

$$P_{y,\tau} v(x) = \chi_{M(y,\tau)}(x) v(x).$$

Then, the boundary data $(\partial M, g_{\partial M})$ and the operator Λ uniquely determine the values $\|P_{y,\tau} u^f(t)\|_{L^2(M)}$ for any $f \in C_0^\infty(\partial M \times \mathbb{R}_+)$, $y \in \partial M$ and $t, \tau > 0$. Let f_ε be of form (44) and (45) and $u_\varepsilon(x,t) = u^f(x,t)$, $f = f_\varepsilon$ be a Gaussian beam propagating along $\gamma_{z_0,\nu}$ described in Lemma 6. The asymptotic expansion (36) of a Gaussian beam implies that for $s < S(z_0)$ and $\tau > 0$,

$$\lim_{\varepsilon \to 0} \|P_{y,\tau} u_\varepsilon(\cdot, s + t_0)\|_{L^2(M)} = \begin{cases} h(s), & \text{for } d_g(\gamma_{z_0,\nu}(s), y) < \tau, \\ 0, & \text{for } d_g(\gamma_{z_0,\nu}(s), y) > \tau, \end{cases} \tag{48}$$

where $h(s)$ is a strictly positive function. By varying $\tau > 0$, we can find $d_g(\gamma_{z_0,\nu}(s), y) = r_x(y)$, where $x = \gamma_{z_0,\nu}(t)$. Moreover, we see that $S(z_0)$ can be determined using the boundary data and (48) by observing that $S(z_0)$ is the smallest number $S > 0$ such that if $t_k \to S$ is an increasing sequence, then

$$d_g(\gamma_{z_0,\nu}(s_k), \partial M) = \inf_{y \in \partial M} d_g(\gamma_{z_0,\nu}(s_k), y) \to 0 \quad \text{as } k \to \infty.$$

Summarizing, for any $z_0 \in \partial M$, we can find $S(z_0)$, and furthermore, for any $0 \le t < S(z_0)$, we can find the boundary distance function $r_x(y)$ with $x = \gamma_{z_0,\nu}(t)$. As

any point $x \in M$ can be represented in this form, we see that the boundary distance representation $R(M)$ can be constructed from the boundary data using the Gaussian beams.

Travel Times and Scattering Relation

We will show in this section that if $(\overline{\Omega}, g)$ is a simple Riemannian manifold, then by looking at the singularities of the response operator, we can determine the boundary distance function $d_g(x, y)$, $x, y \in \partial\Omega$, that is, the travel times of geodesics going through the domain. The boundary distance function is a function of $2n-2$ variables. Thus, the inverse problem of determining the Riemannian metric from the boundary distance function is formally determined in two dimensions and formally overdetermined in dimensions $n \geq 3$.

Let $\Omega \subset \mathbb{R}^n$ be a bounded domain with smooth boundary. If the response operators for the two manifolds $(\overline{\Omega}, g_1)$ and $(\overline{\Omega}, g_2)$ are the same then we can assume, after a change of variables which is the identity at the boundary, the two metrics g_1 and g_2 have the same Taylor series at the boundary [76]. Therefore, we can extend both metrics smoothly to be equal outside Ω and Euclidean outside a ball of radius R. We denote the extensions to $\mathbb{R}^n$ by g_j, $j = 1, 2$, as before. Let $u_j(t, x, \omega)$ be the solution of the continuation problem

$$\begin{cases} \dfrac{\partial^2 u}{\partial t^2} - \Delta_{g_j} u_j = 0, \text{ in } \mathbb{R}^n \times \mathbb{R} \\ \qquad u_j(x, t) = \delta(t - x \cdot \omega), t < -R, \end{cases} \tag{49}$$

where $\omega \in \mathbb{S}^{n-1} = \{x \in \mathbb{R}^n; |x| = 1\}$.

It was shown in [76] that if the response operators for $(\overline{\Omega}, g_1)$ and $(\overline{\Omega}, g_2)$ are equal, then the two solutions coincide outside Ω, namely,

$$u_1(t, x, \omega) = u_2(t, x, \omega), \quad x \in \mathbb{R}^n \backslash \Omega. \tag{50}$$

In the case that the manifold (Ω, g_j), $j = 1, 2$ is simple, we will use methods of geometrical optics to construct solutions of (49) to show that if the response operators of g_1 and g_2 are the same, then the boundary distance functions of the metrics g_1 and g_2 coincide.

Geometrical Optics

Let g denote a smooth Riemannian metric which is Euclidean outside a ball of radius R.

We will construct solutions to the continuation problem for the metric g (which is either g_1 or g_2). We fix ω. Let us assume that there is a solution to Eq. (49) of the form

$$u(x,t,\omega) = a(x,\omega)\delta\big(t-\phi(x,\omega)\big) + v(x,\omega), \quad u = 0, t < -R, \tag{51}$$

where a, ϕ are functions to be determined and $v \in L^2_{loc}$. Notice that in order to satisfy the initial conditions in (49), we require that

$$a = 1, \quad \phi(x,\omega) = x\cdot\omega \text{ for } x\cdot\omega < -R. \tag{52}$$

By replacing Eq. (51) in Eq. (49), it follows that

$$\begin{aligned}\frac{\partial^2 u}{\partial t^2} - \Delta_g u = {} & A\delta''\big(t-\phi(x,\omega)\big) + B\delta'\big(t-\phi(x,\omega)\big) \\ & -(\Delta_g a)\delta\big(t-\phi(x,\omega)\big) + \frac{\partial^2 v}{\partial t^2} - \Delta_g v,\end{aligned} \tag{53}$$

where

$$A = a(x,\omega)\left(1 - \sum_{i,j=1}^{n} g^{ij}\frac{\partial\phi}{\partial x^i}\frac{\partial\phi}{\partial x^j}\right) \tag{54}$$

$$B = 2\sum_{j,k=1}^{n} g^{jk}\frac{\partial a}{\partial x^k}\frac{\partial\phi}{\partial x^j} + a\Delta_g\phi. \tag{55}$$

We choose the functions ϕ, a in the expansion (53) to eliminate the singularities δ'' and δ' and then construct v, so that

$$\frac{\partial^2 v}{\partial t^2} - \Delta_g v = (\Delta_g a)\delta\big(t-\phi(x,\omega)\big), \quad v = 0, t < -R. \tag{56}$$

The Eikonal Equation

In order to solve the equation $A = 0$, it is sufficient to solve the equation

$$\sum_{i,j=1}^{n} g^{ij}\frac{\partial\phi}{\partial x^i}\frac{\partial\phi}{\partial x^j} = 1, \quad \phi(x,\omega) = x\cdot\omega, x\cdot\omega < -R. \tag{57}$$

Equation (57) is known as the *eikonal equation*. Here we will describe a method, using symplectic geometry, to solve this equation.

Let $H_g(x,\xi) = \frac{1}{2}\left(\sum_{i,j=1}^{n} g^{ij}(x)\xi_i\xi_j - 1\right)$ the Hamiltonian associated to the metric g. Note that the metric induced by g in the cotangent space $T^*\mathbb{R}^n$ is given by the principal symbol of the Laplace–Beltrami operator $g^{-1}(x,\xi) = \sum_{i,j=1}^{n} g^{ij}(x)\xi_i\xi_j$. Equation (57) together with the initial condition can be rewritten as

$$H_g(x, d\phi) = 0, \quad \phi(x,\omega) = x\cdot\omega, x\cdot\omega < -R,$$

where $d\phi = \sum_{i=1}^{n} \frac{\partial \phi}{\partial x^i}\, dx^i$ is the differential of ϕ.

Let $S = \{(x,\xi) : H_g(x,\xi) = 0\}$, and let $M_\phi = \{\big(x, \nabla\phi(x)\big) : x \in \mathbb{R}^n\}$, then solving Eq. (57), is equivalent to finding ϕ such that

$$M_\phi \subset S, \text{ with } M_\phi = \{(x,\omega); x\cdot\omega < -R\}. \tag{58}$$

In order to find ϕ so that (58) is valid, we need to find a Lagrangian submanifold L, so that $L \subset S, L = \{(x,\omega); x\cdot\omega < -R\}$ and the projection of $T^*\mathbb{R}^n$ to $\mathbb{R}^n$ is a diffeomorphism [32]. We will construct such a Lagrangian manifold by flowing out from $N = \{(x,\omega) : x\cdot\omega = s \text{ and } s < -R\}$ by the geodesic flow associated to the metric g. We recall the definition of geodesic flow.

We define the Hamiltonian vector field associated to H_g

$$V_g = \left(\frac{\partial H_g}{\partial \xi}, -\frac{\partial H_g}{\partial x}\right). \tag{59}$$

The bicharacteristics are the integral curves of H_g

$$\frac{d}{ds}x^m = \sum_{j=1}^{n} g^{mj}\xi_j, \quad \frac{d}{ds}\xi_m = -\frac{1}{2}\sum_{i,j=1}^{n} \frac{\partial g^{ij}}{\partial x^m}\xi_i\xi_j, m = 1,\dots,n. \tag{60}$$

The projections of the bicharacteristics in the x variable are the geodesics of the metric g and the parameter s denotes arc length. We denote the associated geodesic flow by

$$X_g(s) = \big(x_g(s), \xi_g(s)\big).$$

If we impose the condition that the bicharacteristics are in S initially, then they belong to S for all time, since the Hamiltonian vector field V_g is tangent to S. The Hamiltonian vector field is transverse to N; then the resulting manifold obtained by flowing N along the integral curves of V_g will be a Lagrangian manifold L contained in S. We shall write $L = X_g(N)$.

Now the projection of N into the base space is a diffeomorphism, so that $L = \{(x, d_x\phi)\}$ locally near a point of N. We can construct a global solution of (58) near Ω if the manifold is simple. We recall the definition of simple domains.

Definition 1. Let Ω be a bounded domain of Euclidean space with smooth boundary and g a Riemannian metric on $\overline{\Omega}$. We say that $(\overline{\Omega}, g)$ is simple if given two points on the boundary there is a unique minimizing geodesic joining the two points on the boundary and, moreover, $\partial\Omega$ is geodesically convex.

If $(\overline{\Omega}, g)$ is simple, then we extend the metric smoothly in a small neighborhood, so that the metric g is still simple. In this case we can solve the eikonal equation globally in a neighborhood of Ω.

The Transport Equation

The equation $B = 0$ is equivalent to solving the following equation:

$$\sum_{i,j=1}^{n} g^{ij} \frac{\partial \phi}{\partial x^j} \frac{\partial a}{\partial x^i} + \frac{a}{2} \Delta_g \phi = 0. \tag{61}$$

Equation (61) is called the *transport equation*. It is a vector field equation for $a(x, \omega)$, which is solved by integrating along the integral curves of the vector field $v = \sum_{i,j=1}^{n} g^{ij} \frac{\partial \phi}{\partial x^j} \frac{\partial}{\partial x^i}$. It is an easy computation to prove that v has length 1 and that the integral curves of v are the geodesics of the metric g.

The solution of the transport equation (61) is then given by

$$a(x, \omega) = \exp\left(-\frac{1}{2} \int_{\gamma} \Delta_g \phi\right), \tag{62}$$

where γ is the unique geodesic such that $\gamma(0) = y$, $\dot{\gamma}(0) = \omega$, $y \cdot \omega = 0$ and γ passes through x. If (Ω, g) is a simple manifold, then $a \in C^{\infty}(\mathbb{R}^n)$.

To end the construction of the geometrical optics solutions, we observe that the function $v(t, x, \omega) \in L^2_{\text{loc}}$ by using standard regularity results for hyperbolic equations.

Now we state the main result of this section in the following theorem.

Theorem 6. *Let $(\overline{\Omega}, g_i), i = 1, 2$ be simple manifolds, and assume that the response operators for $(\overline{\Omega}, g_1)$ and $(\overline{\Omega}, g_2)$ are equal. Then $d_{g_1} = d_{g_2}$.*

Sketch of proof 1. Assume that we have two metrics g_1, g_2 with the same response operator. Then by (50), the solutions of (49) are the same outside Ω. Therefore, the main singularity of the solutions in the geometrical optics expansion must be the same outside Ω. Thus, we conclude that

$$\phi_1(x, \omega) = \phi_2(x, \omega), \quad x \in \mathbb{R}^n \backslash \Omega. \tag{63}$$

Now $\phi_j(x, \omega)$ measures the geodesic distance to the hyperplane $x \cdot \omega = -R$ in the metric g. From this, we can easily conclude that the geodesic distance between two points in the boundary for the two metrics is the same, that is, $d_{g_1}(x, y) = d_{g_2}(x, y), x, y \in \partial\Omega$.

This type of argument was used in [61] to study a similar inverse problem for the more complicated system of elastodynamics. In particular, it is proven in [61] that from the response operator associated to the equations of isotropic elastodynamics, one can determine, under the assumption of simplicity of the metrics, the lengths of geodesics of the metrics defined by

$$ds^2 = c_p(x) ds_e^2, \quad ds^2 = c_s(x)^2 ds_e^2, \tag{64}$$

where ds_e is the length element corresponding to the Euclidean metric and $c_p(x) = \sqrt{\frac{(\lambda+2\mu)}{\rho}}$, $c_s(x) = \sqrt{\frac{\mu}{\rho}}$ denote the speed of compressional waves and shear waves, respectively. Here λ, μ are the Lamé parameters and ρ the density.

Using Mukhometov's result [56, 57], we can recover both speeds from the response operator. This shows in particular that if we know the density, one can determine the Lamé parameters from the response operator. By using the transport equation of geometrical optics, similar to (61), and the results on the ray transform (see, e.g., [66]), Rachele shows that under certain a priori conditions one can also determine the density ρ [62].

Scattering Relation

In the presence of caustics (i.e., the exponential map is not a diffeomorphism), the expansion (51) is not valid since we cannot solve the eikonal equation globally in Ω. The solution of (50) is globally a Lagrangian distribution (see, e.g., [38]). These distributions can locally be written in the form

$$u(t,x,\omega) = \int_{\mathbb{R}^m} e^{i\phi(t,x,\omega,\theta)} a(t,x,\omega,\theta)\, d\theta, \tag{65}$$

where ϕ is a phase function and $a(t,x,\omega)$ is a classical symbol.

Every Lagrangian distribution is determined (up to smoother terms) by a Lagrangian manifold and its symbol. The Lagrangian manifold associated to $u(t,x,\omega)$ is the flow out from $t = x\cdot\omega, t < -R$ by the Hamilton vector field of $p_g(t,x,\tau,\xi) = \tau^2 - \sum_{j,k=1}^n g_{jk}(x)\xi^j\xi^k$. Here (τ,ξ) are the dual variables to (t,x), respectively. The projection in the (x,ξ) variables of the flow is given by the flow out from N by geodesic flow, that is, the Lagrangian submanifold L described above.

The *scattering relation* (also called lens map) $C_g \subset \big(T^*(\mathbb{R}\times\partial\Omega)\backslash 0\big)\times\big(T^*(\mathbb{R}\times\partial\Omega)\backslash 0\big)$ of a metric $g = (g^{ij})$ on $\overline{\Omega}$ with dual metric $g^{-1} = (g_{ij})$ is defined as follows. Consider bicharacteristic curves, $\gamma : [a,b] \to T^*(\overline{\Omega}\times\mathbb{R})$, of the Hamilton function $p_g(t,x,\tau,\xi)$, which satisfy the following: $\gamma(]a,b[)$ lies in the interior, γ intersects the boundary non-tangentially at $\gamma(a)$ and $\gamma(b)$, and time increases along γ. Then the canonical projection from $\big(T^*_{\mathbb{R}\times\partial\Omega}(\mathbb{R}\times\Omega)\backslash 0\big)\times\big(T^*_{\mathbb{R}\times\partial\Omega}(\mathbb{R}\times\Omega)\backslash 0\big)$ onto $\big(T^*(\mathbb{R}\times\partial\Omega)\backslash 0\big)\times T^*(\mathbb{R}\times\partial\Omega)\backslash 0\big)$ maps the endpoint pair $\big(\gamma(b),\gamma(a)\big)$ to a point in C_g. In other words, C_g gives the geodesic distance between points in the boundary and also the points of exit and direction of exit of the geodesic if we know the point of entrance and direction of entrance.

It is well known that C_g is a homogeneous canonical relation on $\big(\big(T^*(\mathbb{R}\times\partial\Omega)\backslash 0\big)\times\big(T^*(\mathbb{R}\times\partial\Omega)\backslash 0\big)$. (See [35] for the concept of a scattering relation.) C_g is, in fact, a diffeomorphism between open subsets of $T^*(\mathbb{R}\times\partial\Omega)\backslash 0$.

In analogy with Theorem 6, we have the following theorem.

Theorem 7. *Let $g_i, i = 1,2$ be Riemannian metrics on $\overline{\Omega}$ such that the response operators for $(\overline{\Omega}, g_1)$ and $(\overline{\Omega}, g_2)$ are equal. Then*

$$C_{g_1} = C_{g_2}.$$

Sketch of proof 2. Since by (49), we know the solutions of (48) outside Ω. Therefore, the associated Lagrangian manifolds to the Lagrangian distributions u_j must be the same outside Ω. By taking the projection of these Lagrangians onto the boundary, we get the desired claim.

In the case that $(\overline{\Omega}, g)$ is simple, the scattering relation does not give any new information. In fact $\big((t_1, x_1, \tau, \xi_1), (t_0, x_0, \tau, \xi_0)\big) \in C_g$ if $t_1 - t_0 = d_g(x_1, x_0)$ and $\xi_j = -\tau \frac{\partial d_g(x_1, x_0)}{\partial x^j}$, $j = 0, 1$. In other words d_g is the generating function of the scattering relation.

This result was generalized in [36] to the case of the equations of elastodynamics with residual stress. It is shown that knowing the response operator, we can recover the scattering relations associated to P and S waves. For this, one uses Lagrangian distributions with appropriate polarization.

The scattering relation contains all travel time data; not just information about minimizing geodesics as is the case of the boundary distance function. The natural conjecture is that on a nontrapping manifold, this is enough to determine the metric up to isometry. We refer to [72] and the references therein for results on this problem.

Curvelets and Wave Equations

In this section we will discuss in more detail the use of curvelets in wave imaging. We begin by explaining the curvelet decomposition of functions, using the standard second dyadic decomposition of phase space. The curvelets provide tight frames of $L^2(\mathbb{R}^n)$ and give efficient representations of sharp wave fronts. We then discuss why curvelets are useful for solving the wave equation. This is best illustrated in terms of the half-wave equation (a first-order hyperbolic equation), where a good approximation to the solution is obtained by decomposing the initial data in curvelets and then by translating each curvelet along the Hamilton flow for the equation. Then we explain how one deals with wave speeds of limited smoothness, and how one can convert the approximate solution operator into an exact one by doing a Volterra iteration.

The treatment below follows the original approach of Smith [67] and focuses on explaining the theoretical aspects of curvelet methods for solving wave equations. We refer to the works mentioned in the introduction for applications and more practical considerations.

Curvelet Decomposition

We will explain the curvelet decomposition in its most standard form, as given in [67]. In a nutshell, curvelets are functions which are frequency localized in certain frequency shells and certain directions, according to the second dyadic

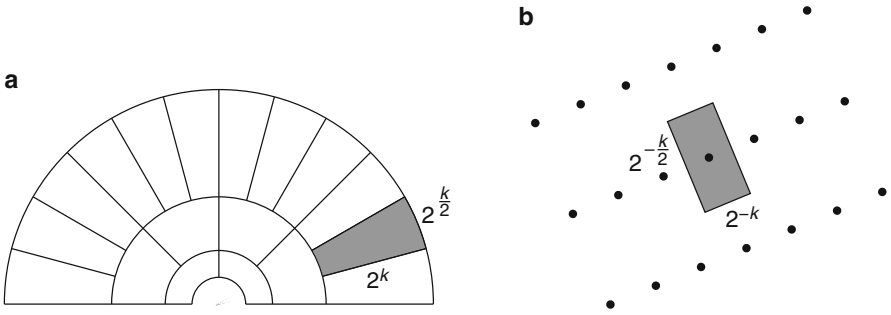


Fig. 3 A curvelet φ_γ with $\gamma = (k, \omega, x)$ is concentrated (**a**) in the frequency domain near a box of length $\sim 2^k$ and width $\sim 2^{k/2}$ and (**b**) in the spatial side near a box of length $\sim 2^{-k}$ and width $\sim 2^{-k/2}$

decomposition and parabolic scaling. On the spatial side, curvelets are concentrated near lattice points which correspond to the frequency localization.

To make this more precise, we recall the *dyadic decomposition* of the frequency space $\{\xi \in \mathbb{R}^n\}$ into the ball $\{|\xi| \leq 1\}$ and dyadic shells $\{2^k \leq |\xi| \leq 2^{k+1}\}$. The *second dyadic decomposition* further subdivides each frequency shell $\{2^k \leq |\xi| \leq 2^{k+1}\}$ into slightly overlapping "boxes" of width roughly $2^{k/2}$ (thus each box resembles a rectangle whose major axis has length $\sim 2^k$ and all other axes have length $\sim 2^{k/2}$). See Fig. 3a for an illustration. The convention that the width $(2^{k/2})$ of the boxes is the square root of the length (2^k) is called *parabolic scaling*; this scaling is crucial for the wave equation as will be explained later.

In the end, the second dyadic decomposition amounts to having a collection of nonnegative functions $h_0, h_k^\omega \in C_c^\infty(\mathbb{R}^n)$, which form a partition of unity in the sense that

$$1 = h_0(\xi)^2 + \sum_{k=0}^{\infty} \sum_{\omega} h_k^\omega(\xi)^2.$$

Here, for each k, ω runs over roughly $2^{(n-1)k/2}$ unit vectors uniformly distributed over the unit sphere, and h_k^ω is supported in the set

$$2^{k-1/2} \leq |\xi| \leq 2^{k+3/2}, \quad \left| \frac{\xi}{|\xi|} - \omega \right| \leq 2^{-k/2}.$$

We also require a technical estimate for the derivatives

$$|\langle \omega, \partial_\xi \rangle^j \partial_\xi^\alpha h_k^\omega(\xi)| \leq C_{j,\alpha} 2^{-k(j+|\alpha|/2)},$$

with $C_{j,\alpha}$ independent of k and ω. Such a partition of unity is not hard to construct; we refer to [73, Sect. 20.9.4] for the details.

On the frequency side, a curvelet at frequency level 2^k with direction ω will be supported in a rectangle with side length $\sim 2^k$ in direction ω and side lengths $\sim 2^{k/2}$ in the orthogonal directions. By the uncertainty principle, on the spatial side, one expects a curvelet to be concentrated in a rectangle with side length $\sim 2^{-k}$ in direction ω and $\sim 2^{-k/2}$ in other directions. Motivated by this, we define a rectangular lattice Ξ_k^ω in $\mathbb{R}^n$, which has spacing 2^{-k} in direction ω and spacing $2^{-k/2}$ in the orthogonal directions, thus

$$\Xi_k^\omega = \left\{ x \in \mathbb{R}^n \,;\, x = a2^{-k}\omega + \sum_{j=2}^{n} b_j 2^{-k/2}\omega_j \text{ where } a, b_j \in \mathbb{Z} \right\}$$

and $\{\omega, \omega_2, \ldots, \omega_n\}$ is a fixed orthonormal basis of $\mathbb{R}^n$. See Fig. 3b.

We are now ready to give a definition of the curvelet frame.

Definition 2. For a triplet $\gamma = (k, \omega, x)$ with ω as described above and for $x \in \Xi_k^\omega$, we define the corresponding *fine-scale curvelet* φ_γ in terms of its Fourier transform by

$$\hat{\varphi}_\gamma(\xi) = (2\pi)^{-n/2} 2^{-k(n+1)/4} e^{-ix\cdot\xi} h_k^\omega(\xi).$$

The *coarse-scale curvelets* for $\gamma = (0, x)$ with $x \in \mathbb{Z}^n$ are given by

$$\hat{\varphi}_\gamma(\xi) = (2\pi)^{-n/2} e^{-ix\cdot\xi} h_0(\xi).$$

The distinction between coarse- and fine-scale curvelets is analogous to the case of wavelets. The coarse-scale curvelets are used to represent data at low frequencies $\{|\xi| \le 1\}$, and they are direction independent, whereas the fine-scale curvelets depend on the direction ω.

The next list collects some properties of the (fine-scale) curvelets φ_γ.

- *Frequency localization.* The Fourier transform $\hat{\varphi}_\gamma(\xi)$ is supported in the shell $\{2^{k-1/2} < |\xi| < 2^{k+3/2}\}$ and in a rectangle with side length $\sim 2^k$ in the ω direction and side length $\sim 2^{k/2}$ in directions orthogonal to ω.
- *Spatial localization.* The function $\varphi_\gamma(y)$ is concentrated in (i.e., decays away from) a rectangle centered at $x \in \Xi_k^\omega$, having side length 2^{-k} in the ω direction and side lengths $2^{-k/2}$ in directions orthogonal to ω.
- *Tight frame.* Any function $f \in L^2(\mathbb{R}^n)$ may be written in terms of curvelets as

$$f(y) = \sum_\gamma c_\gamma \varphi_\gamma(y),$$

where c_γ are the curvelet coefficients of f:

$$c_\gamma = \int_{\mathbb{R}^n} f(y)\overline{\varphi_\gamma(y)}\, dy.$$

One has the Plancherel identity

$$\int_{\mathbb{R}^n} |f(y)|^2\, dy = \sum_\gamma |c_\gamma|^2.$$

The last statement about how to represent a function $f \in L^2(\mathbb{R}^n)$ in terms of curvelets can be proved by writing

$$\hat{f}(\xi) = h_0(\xi)^2 \hat{f}(\xi) + \sum_{k=0}^{\infty} \sum_{\omega} h_k^\omega(\xi)^2 \hat{f}(\xi)$$

and then by expanding the functions $h_k^\omega(\xi)\hat{f}(\xi)$ in Fourier series in suitable rectangles, and finally by taking the inverse Fourier transform. Note that any L^2 function can be represented as a superposition of curvelets φ_γ, but that the φ_γ are not orthogonal and the representation is not unique.

Curvelets and Wave Equations

Next we explain, in a purely formal way, how curvelets can be used to solve the Cauchy problem for the wave equation

$$\begin{aligned} \left(\partial_t^2 + A(x, D_x)\right) u(t,x) &= F(t,x) \quad \text{in } \mathbb{R} \times \mathbb{R}^n, \\ u(0,x) &= u_0(x), \\ \partial_t u(0,x) &= u_1(x). \end{aligned}$$

Further details and references are given in the next section. Here $A(x, D_x) = \sum_{j,k=1}^n g^{jk}(x) D_{x_j} D_{x_k}$ is a uniform elliptic operator, meaning that $g^{jk} = g^{kj}$ and $0 < \lambda \le \sum_{j,k=1}^n g^{jk}(x)\xi_j\xi_k \le \Lambda < \infty$ uniformly over $x \in \mathbb{R}^n$ and $\xi \in S^{n-1}$. We assume that g^{jk} are smooth and have uniformly bounded derivatives of all orders.

It is enough to construct an operator $S(t) : u_1 \mapsto u(t, \cdot)$ such that $u(t,x) = (S(t)u_1)(x)$ solves the above wave equation with $F \equiv 0$ and $u_0 \equiv 0$. Then, by Duhamel's principle, the general solution of the above equation will be

$$u(t,x) = \int_0^t S(t-s)F(s,x)\, ds + \left(\partial_t S(t)u_0\right)(x) + \left(S(t)u_1\right)(x).$$

To construct $S(t)$, we begin by factoring the wave operator $\partial_t^2 + A(x, D_x)$ into two first-order hyperbolic operators, known as *half-wave operators*. Let $P(x, D_x) = \sqrt{A(x, D_x)}$ be a formal square root of the elliptic operator $A(x, D_x)$. Then we have

$$\partial_t^2 + A(x, D_x) = (\partial_t - iP)(\partial_t + iP)$$

and the Cauchy problem for the wave equation with data $F \equiv 0, u_0 \equiv 0, u_1 = f$ is reduced to solving the two first-order equations

$$\begin{aligned}(\partial_t - iP)v &= 0, \quad v(0) = f,\\ (\partial_t + iP)u &= v, \quad u(0) = 0.\end{aligned}$$

If one can solve the first equation, then solvability of the second equation will follow from Duhamel's principle (the sign in front of P is immaterial).

Therefore, we only need to solve

$$\begin{aligned}(\partial_t - iP)v(t, x) &= 0,\\ v(0, x) &= f(x).\end{aligned}$$

For the moment, let us simplify even further and assume that $A(x, D_x)$ is the Laplacian $-\Delta$, so that P will be the operator given by

$$\widehat{Pf}(\xi) = |\xi| \hat{f}(\xi).$$

Taking the spatial Fourier transform of the equation for v and solving the resulting ordinary differential equation give the full solution

$$v(t, y) = (2\pi)^{-n} \int_{\mathbb{R}^n} e^{i(y\cdot\xi + t|\xi|)} \hat{f}(\xi)\, d\xi.$$

Thus, the solution is given by a Fourier integral operator acting on f:

$$v(t, y) = (2\pi)^{-n} \int_{\mathbb{R}^n} e^{i\Phi(t,y,\xi)} a(t, y, \xi) \hat{f}(\xi)\, d\xi.$$

In this particular case, the phase function is $\Phi(t, y, \xi) = y \cdot \xi + t|\xi|$, and the symbol is $a(t, y, \xi) \equiv 1$.

So far we have not used any special properties of f. Here comes the key point. *If f is a curvelet, then the phase function is well approximated on* $\operatorname{supp}(f)$ *by its linearization in ξ:*

$$\Phi(t, y, \xi) \approx \nabla_\xi \Phi(t, y, \omega) \cdot \xi \quad \text{for } \xi \in \operatorname{supp}(f).$$

(This statement may seem somewhat mysterious, but it really is one reason why curvelets are useful for wave imaging. A slightly more precise statement is as follows: if $\Psi(t, y, \xi)$ is smooth for $\xi \neq 0$, homogeneous of order 1 in ξ, and its derivatives are uniformly bounded over $t \in [-T, T]$ and $y \in \mathbb{R}^n$ and $\xi \in S^{n-1}$, then

$$|\Psi(t,y,\xi) - \nabla_\xi \Psi(t,y,\omega)\cdot\xi| \lesssim 1$$

whenever $\xi\cdot\omega \sim 2^k$ and $|\xi - (\xi\cdot\omega)\omega| \lesssim 2^{k/2}$. Also the derivatives of $\Psi(t,y,\xi) - \nabla_\xi\Psi(t,y,\omega)\cdot\xi$ satisfy suitable symbol bounds. Parabolic scaling is crucial here; we refer to [18, section "Travel Times and Scattering Relation"] for more on this point.) Thus, if $f = \varphi_\gamma$ then the solution v with this initial data is approximately given by

$$v(t,y) \approx (2\pi)^{-n}\int_{\mathbb{R}^n} e^{i(y+t\omega)\cdot\xi}\hat{\varphi}_\gamma(\xi)\,d\xi = \varphi_\gamma(y+t\omega).$$

Thus the half-wave equation for $P = \sqrt{-\Delta}$, whose initial data is a curvelet in direction ω, is approximately solved by translating the curvelet along a straight line in direction ω.

We now return to the general case, where $A(x,\xi)$ is a general elliptic symbol $\sum_{j,k=1}^n g^{jk}(x)\xi_j\xi_k$. We define

$$p(x,\xi) = \sqrt{A(x,\xi)}.$$

Then p is homogeneous of order 1 in ξ, and it generates a *Hamilton flow* $\big(x(t),\xi(t)\big)$ in the phase space $T^*\mathbb{R}^n = \mathbb{R}^n\times\mathbb{R}^n$, determined by the ordinary differential equations

$$\begin{aligned}\dot{x}(t) &= \nabla_\xi p\big(x(t),\xi(t)\big),\\ \dot{\xi}(t) &= -\nabla_x p\big(x(t),\xi(t)\big).\end{aligned}$$

If $A(x,\xi)$ is smooth, then the curves $\big(x(t),\xi(t)\big)$ starting at some point $\big(x(0),\xi(0)\big) = (x,\omega)$ are smooth and exist for all time. Note that if $p(x,\xi) = |\xi|$, then one has straight lines $\big(x(t),\xi(t)\big) = (x+t\omega,\omega)$.

Similarly as above, the half-wave equation

$$\begin{aligned}(\partial_t - iP)v(t,x) &= 0,\\ v(0,x) &= f(x)\end{aligned}$$

can be approximately solved as follows:

1. Write the initial data f in terms of curvelets as $f(y) = \sum_\gamma c_\gamma\varphi_\gamma(y)$.
2. For a curvelet $\varphi_\gamma(y)$ centered at x pointing in direction ω, let $\varphi_\gamma(t,y)$ be another curvelet centered at $x(t)$ pointing in direction $\xi(t)$. That is, translate each curvelet φ_γ for time t along the Hamilton flow for P.
3. Let $v(t,y) = \sum_\gamma c_\gamma\varphi_\gamma(t,y)$ be the approximate solution.

Thus, the wave equation can be approximately solved by decomposing the initial data into curvelets and then by translating each curvelet along the Hamilton flow.

Low-Regularity Wave Speeds and Volterra Iteration

Here we give some further details related to the formal discussion in the previous section, following the arguments in [67]. The precise assumption on the coefficients will be

$$g^{jk}(x) \in C^{1,1}(\mathbb{R}^n).$$

This means that $\partial^\alpha g^{jk} \in L^\infty(\mathbb{R}^n)$ for $|\alpha| \le 2$, which is a minimal assumption which guarantees a well-defined Hamilton flow.

As discussed in section "Curvelets and Wave Equations," by Duhamel's formula, it is sufficient to consider the Cauchy problem

$$\begin{aligned} \left(\partial_t^2 + A(x, D_x)\right) u(t,x) &= 0 \quad \text{in } \mathbb{R} \times \mathbb{R}^n, \\ u(0,x) &= 0, \\ \partial_t u(0,x) &= f. \end{aligned}$$

Here, $A(x, D_x) = \sum_{j,k=1}^n g^{jk}(x) D_{x_j} D_{x_k}$ and $g^{jk} \in C^{1,1}(\mathbb{R}^n)$, $g^{jk} = g^{kj}$, and $0 < \lambda \le \sum_{j,k=1}^n g^{jk}(x)\xi_j\xi_k \le \Lambda < \infty$ uniformly over $x \in \mathbb{R}^n$ and $\xi \in S^{n-1}$.

To deal with the nonsmooth coefficients, we introduce the smooth approximations

$$A_k(x,\xi) = \sum_{i,j=1}^n g_k^{ij}(x)\xi_i\xi_j, \quad g_k^{ij} = \chi(2^{-k/2}D_x)g^{ij}$$

where $\chi \in C_c^\infty(\mathbb{R}^n)$ satisfies $0 \le \chi \le 1$, $\chi(\xi) = 1$ for $|\xi| \le 1/2$, and $\chi(\xi) = 0$ for $|\xi| \ge 1$. We have written $\left(\chi(2^{-k/2}D_x)g\right)\hat{}\,(\xi) = \chi(2^{-k/2}\xi)\hat{g}(\xi)$. Thus g_k^{ij} are smooth truncations of g^{ij} to frequencies $\le 2^{k/2}$. We will use the smooth approximation A_k in the construction of the solution operator at frequency level 2^k, which is in keeping with paradifferential calculus.

Given a curvelet $\varphi_\gamma(y)$ where $\gamma = (k, \omega_\gamma, x_\gamma)$, we wish to consider a curvelet $\varphi_\gamma(t,y)$ which corresponds to a translation of φ_γ for time t along the Hamilton flow for $H_k(x,\xi) = \sqrt{A_k(x,\xi)}$. In fact, we shall define

$$\varphi_\gamma(t,y) = \varphi_\gamma\left(\Theta_\gamma(t)\left(y - x_\gamma(t)\right) + x_\gamma\right),$$

where $x_\gamma(t)$ and the $n \times n$ matrix $\Theta_\gamma(t)$ arise as the solution of the equations

$$\begin{aligned} \dot{x} &= \nabla_\xi H_k(x,\omega), \\ \dot{\omega} &= -\nabla_x H_k(x,\omega) + \left(\omega \cdot \nabla_x H_k(x,\omega)\right)\omega, \end{aligned}$$

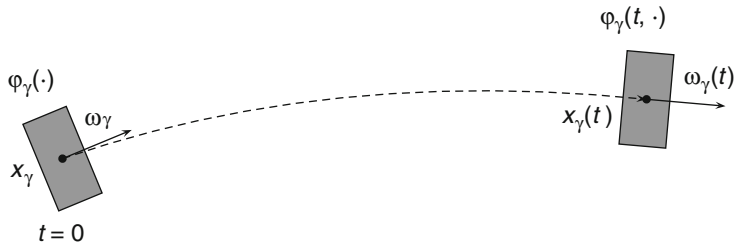


Fig. 4 The translation of a curvelet φ_γ for time t along the Hamilton flow

$$\dot{\Theta} = -\Theta\left(\omega \otimes \nabla_x H_k(x,\omega) - \nabla_x H_k(x,\omega) \otimes \omega\right)$$

with initial condition $\big(x_\gamma(0), \omega_\gamma(0), \Theta_\gamma(0)\big) = (x_\gamma, \omega_\gamma, I)$. Here $v \otimes w$ is the matrix with $(v \otimes w)x = (w \cdot x)v$. The idea is that $\big(x_\gamma(t), \omega_\gamma(t)\big)$ is the Hamilton flow for H_k restricted to the unit cosphere bundle $S^*\mathbb{R}^n = \{(x,\xi) \in T^*\mathbb{R}^n\,;\, |\xi| = 1\}$, and $\Theta_\gamma(t)$ is a matrix which tracks the rotation of ω_γ along the flow and satisfies $\Theta_\gamma(t)\omega_\gamma(t) = \omega_\gamma$ for all t. See Fig. 4 for an illustration.

We define an approximate solution operator at frequency level 2^k by

$$E_k(t)f(y) = \sum_{\gamma' : k' = k} (f, \varphi_{\gamma'})_{L^2(\mathbb{R}^n)} \varphi_{\gamma'}(t, y).$$

Summing over all frequencies, we consider the operator

$$E(t)f = \sum_{k=0}^{\infty} E_k(t)f.$$

This operator essentially takes a function f, decomposes it into curvelets, and then translates each curvelet at frequency level 2^k for time t along the Hamilton flow for H_k.

It is proved in [67, Theorem 3.2] that $E(t)$ is an operator of order 0, mapping $H^\alpha(\mathbb{R}^n)$ to $H^\alpha(\mathbb{R}^n)$ for any α. The fact that $E(t)$ is an approximate solution operator is encoded in the result that the wave operator applied to $E(t)$,

$$T(t) = \left(\partial_t^2 + A(x, D_x)\right) E(t),$$

which is a priori a second-order operator, is in fact an operator of order 1 and maps $H^{\alpha+1}(\mathbb{R}^n)$ to $H^\alpha(\mathbb{R}^n)$ for $-1 \le \alpha \le 2$. This is proved in [67, Theorem 4.5] and is due to the two facts. The first one is that when A is replaced by the smooth approximation A_k, the corresponding operator

$$\sum_k \left(\partial_t^2 + A_k(x, D_x)\right) E_k(t)$$

is of order 1 because the second-order terms cancel. Here, one uses that translation along the Hamilton flow approximately solves the wave equation. The second fact is that the part involving the nonsmooth coefficients

$$\sum_k \left(A_k(x, D_x) - A(x, D_x)\right) E_k(t)$$

is also of order 1 using that A_k is truncated to frequencies $\leq 2^{k/2}$ and using estimates for $A - A_k$ obtained from the $C^{1,1}$ regularity of the coefficients.

To obtain the full parametrix, one needs to consider the Hamilton flows both for $\sqrt{A_k}$ and $-\sqrt{A_k}$, corresponding to the two half-wave equations appearing in the factorization of the wave operator, and one also needs to introduce corrections to ensure that the initial values of the approximate solution are the given functions. For simplicity, we will not consider these details here and only refer to [67, Sect. 4]. The outcome of this argument is an operator $\mathbf{s}(t, s)$, which is strongly continuous in t and s as a bounded operator $H^\alpha(\mathbb{R}^n) \to H^{\alpha+1}(\mathbb{R}^n)$ satisfies $\mathbf{s}(t, s) f|_{t=s} = 0$ and $\partial_t \mathbf{s}(t, s) f|_{t=s} = f$, and further the operator

$$T(t, s) = \left(\partial_t^2 + A(x, D_x)\right) \mathbf{s}(t, s)$$

is bounded $H^\alpha(\mathbb{R}^n) \to H^\alpha(\mathbb{R}^n)$ for $-1 \leq \alpha \leq 2$.

We conclude this discussion by explaining the Volterra iteration scheme, which is used for converting the approximate solution operator to an exact one, as in [67, Theorem 4.6]. We look for a solution in the form

$$u(t) = \mathbf{s}(t, 0) f + \int_0^t \mathbf{s}(t, s) G(s)\, ds$$

for some $G \in L^1\left([-t_0, t_0]\,;\, H^\alpha(\mathbb{R}^n)\right)$. From the properties of $\mathbf{s}(t, s)$, we see that u satisfies

$$\left(\partial_t^2 + A(x, D_x)\right) u = T(t, 0) f + G(t) + \int_0^t T(t, s) G(s)\, ds.$$

Thus, u is a solution if G is such that

$$G(t) + \int_0^t T(t, s) G(s)\, ds = -T(t, 0) f.$$

Since $T(t, s)$ is bounded on $H^\alpha(\mathbb{R}^n)$ for $-1 \leq \alpha \leq 2$, with norm bounded by a uniform constant when $|t|, |s| \leq t_0$, the last Volterra equation can be solved by iteration. This yields the required solution u.

4 Conclusion

In this chapter, inverse problems for the wave equation were considered with different types of data. All considered data correspond to measurements made on the boundary of a body in which the wave speed is unknown and possibly anisotropic. The case of the complete data, that is, with measurements of amplitudes and phases of waves corresponding to all possible sources on the boundary, was considered using the boundary control method. We showed that the wave speed can be reconstructed from the boundary measurements up to a diffeomorphism of the domain. This corresponds to the determination of the wave speed in the local travel time coordinates. Next, the inverse problem with less data, the scattering relation, was considered. The scattering relation consists of the travel times and the exit directions of the wave fronts produced by the point sources located on the boundary of the body. Such data can be considered to be obtained by measuring the waves up to smooth errors or measuring only the singularities of the waves. The scattering relation is a generalization of the travel time data, that is, the travel times of the waves through the body. Finally, we considered the use of wavelets and curvelets in the analysis of the waves. Using the curvelet representation of the waves, the singularities of the waves can be efficiently analyzed. In particular, the curvelets are suitable for the simulation of the scattering relation, even when the wave speed is nonsmooth. Summarizing, in this chapter, modern approaches to study inverse problems for wave equations based on the control theory, the geometry, and the microlocal analysis were presented.

Cross-References

▸ Inverse Scattering
▸ Photoacoustic and Thermoacoustic Tomography: Image Formation Principles

References

1. Anderson, M., Katsuda, A., Kurylev, Y., Lassas, M., Taylor, M.: Boundary regularity for the Ricci equation, geometric convergence, and Gel'fand's inverse boundary problem. Invent. Math. **158**, 261–321 (2004)
2. Andersson, F., de Hoop, M.V., Smith, H.F., Uhlmann, G.: A multi-scale approach to hyperbolic evolution equations with limited smoothness. Commun. Part. Differ. Equ. **33**(4–6), 988–1017 (2008)
3. Babich, V.M., Ulin, V.V.: The complex space-time ray method and "quasiphotons" (Russian). Zap. Nauchn. Sem. LOMI **117**, 5–12 (1981)
4. Belishev, M.: An approach to multidimensional inverse problems for the wave equation (Russian). Dokl. Akad. Nauk SSSR **297**(3), 524–527 (1987)
5. Belishev, M.: Boundary control in reconstruction of manifolds and metrics (the BC method). Inverse Probl. **13**, R1–R45 (1997)

6. Belishev, M., Kachalov, A.: Boundary control and quasiphotons in a problem of the reconstruction of a Riemannian manifold from dynamic data (Russian). Zap. Nauchn. Sem. POMI **203**, 21–50 (1992)
7. Belishev, M., Kurylev, Y.: To the reconstruction of a Riemannian manifold via its spectral data (BC-method). Commun. Part. Differ. Equ. **17**, 767–804 (1992)
8. Bernstein, I.N., Gerver, M.L.: Conditions on Distinguishability of Metrics by Hodographs, Methods and Algorithms of Interpretation of Seismological Information. Computerized Seismology, vol. 13, pp. 50–73. Nauka, Moscow (1980) (in Russian)
9. Besson, G., Courtois, G., Gallot, S.: Entropies et rigidités des espaces localement symétriques de courbure strictment négative. Geom. Funct. Anal. **5**, 731–799 (1995)
10. Beylkin, G.: Stability and uniqueness of the solution of the inverse kinematic problem in the multidimensional case. J. Sov. Math. **21**, 251–254 (1983)
11. Bingham, K., Kurylev, Y., Lassas, M., Siltanen, S.: Iterative time reversal control for inverse problems. Inverse Probl. Imaging **2**, 63–81 (2008)
12. Blagoveščenskii, A.: A one-dimensional inverse boundary value problem for a second order hyperbolic equation (Russian). Zap. Nauchn. Sem. LOMI **15**, 85–90 (1969)
13. Blagoveščenskii, A.: Inverse boundary problem for the wave propagation in an anisotropic medium (Russian). Trudy Mat. Inst. Steklova **65**, 39–56 (1971)
14. Brytik, V., de Hoop, M.V., Salo, M.: Sensitivity analysis of wave-equation tomography: a multiscale approach. J. Fourier Anal. Appl. **16**(4), 544–589 (2010)
15. Burago, D., Ivanov, S.: Boundary rigidity and filling volume minimality of metrics close to a flat one. Ann. Math. **171**(2), 1183–1211 (2010)
16. Burago, D., Ivanov, S.: Area minimizers and boundary rigidity of almost hyperbolic metrics (in preparation)
17. Candès, E.J., Demanet, L.: Curvelets and Fourier integral operators. C. R. Math. Acad. Sci. Paris **336**, 395–398 (2003)
18. Candès, E.J., Demanet, L.: The curvelet representation of wave propagators is optimally sparse. Commun. Pure Appl. Math. **58**, 1472–1528 (2005)
19. Candès, E.J., Donoho, D.L.: Curvelets – a surprisingly effective nonadaptive representation for objects with edges. In: Schumaker, L.L., et al. (eds.) Curves and Surfaces, pp. 105–120. Vanderbilt University Press, Nashville (2000)
20. Candès, E.J., Donoho, D.L.: New tight frames of curvelets and optimal representations of objects with piecewise C2 singularities. Commun. Pure Appl. Math. **57**, 219–266 (2004)
21. Candès, E.J., Demanet, L., Ying, L.: Fast computation of Fourier integral operators. SIAM J. Sci. Comput. **29**, 2464–2493 (2007)
22. Chavel, I.: Riemannian Geometry. A Modern Introduction, pp. xvi+471. Cambridge University Press, Cambridge (2006)
23. Córdoba, A., Fefferman, C.: Wave packets and Fourier integral operators. Commun. Part. Differ. Equ. **3**, 979–1005 (1978)
24. Creager, K.C.: Anisotropy of the inner core from differential travel times of the phases PKP and PKIPK. Nature **356**, 309–314 (1992)
25. Croke, C.: Rigidity for surfaces of non-positive curvature. Comment. Math. Helv. **65**, 150–169 (1990)
26. Croke, C.: Rigidity and the distance between boundary points. J. Differ. Geom. **33**(2), 445–464 (1991)
27. Dahl, M., Kirpichnikova, A., Lassas, M.: Focusing waves in unknown media by modified time reversal iteration. SIAM J. Control Optim. **48**, 839–858 (2009)
28. de Hoop, M.V.: Microlocal analysis of seismic inverse scattering: inside out. In: Uhlmann, G. (ed.) Inverse Problems and Applications, pp. 219–296. Cambridge University Press, Cambridge (2003)
29. de Hoop, M.V., Smith, H., Uhlmann, G., van der Hilst, R.D.: Seismic imaging with the generalized Radon transform: a curvelet transform perspective. Inverse Probl. **25**(2), 25005–25025 (2009)

30. Demanet, L., Ying, L.: Wave atoms and time upscaling of wave equations. Numer. Math. **113**(1), 1–71 (2009)
31. Duchkov, A.A., Andersson, F., de Hoop, M.V.: Discrete, almost symmetric wave packets and multiscale geometric representation of (seismic) waves. IEEE Trans. Geosci. Remote Sens. **48**(9), 3408–3423 (2010)
32. Duistermaat, J.J.: Fourier Integral Operators. Birkhäuser, Boston (2009)
33. Greenleaf, A., Kurylev, Y., Lassas, M., Uhlmann, G.: Invisibility and inverse problems. Bull. Am. Math. **46**, 55–97 (2009)
34. Gromov, M.: Filling Riemannian manifolds. J. Differ. Geom. **18**(1), 1–148 (1983)
35. Guillemin, V.: Sojourn times and asymptotic properties of the scattering Matrix. In: Proceedings of the Oji Seminar on Algebraic Analysis and the RIMS Symposium on Algebraic Analysis, Kyoto. Kyoto University, Kyoto (1976). Publ. Res. Inst. Math. Sci. **12**(1976/77, Suppl), 69–88
36. Hansen, S., Uhlmann, G.: Propagation of polarization for the equations in elastodynamics with residual stress and travel times. Math. Ann. **326**, 536–587 (2003)
37. Herglotz, G.: Uber die Elastizitaet der Erde bei Beruecksichtigung ihrer variablen Dichte. Zeitschr. fur Math. Phys. **52**, 275–299 (1905)
38. Hörmander, L.: The Analysis of Linear Partial Differential Operators III. Pseudodifferential Operators, pp. viii+525. Springer, Berlin (1985)
39. Isozaki, H., Kurylev, Y., Lassas, M.: Forward and Inverse scattering on manifolds with asymptotically cylindrical ends. J. Funct. Anal. **258**, 2060–2118 (2010)
40. Ivanov, S.: Volume comparison via boundary distances. arXiv:1004–2505
41. Katchalov, A., Kurylev, Y.: Multidimensional inverse problem with incomplete boundary spectral data. Commun. Part. Differ. Equ. **23**, 55–95 (1998)
42. Katchalov, A., Kurylev, Y., Lassas, M.: Inverse Boundary Spectral Problems, pp. xx+290. Chapman & Hall/CRC, Boca Raton (2001)
43. Katchalov, A., Kurylev, Y., Lassas, M.: Energy measurements and equivalence of boundary data for inverse problems on non-compact manifolds. In: Croke, C., Lasiecka, I., Uhlmann, G., Vogelius, M. (eds.) Geometric Methods in Inverse Problems and PDE Control. IMA Volumes in Mathematics and Applications, vol. 137, pp. 183–213. Springer, New York (2004)
44. Katchalov, A., Kurylev, Y., Lassas, M., Mandache, N.: Equivalence of time-domain inverse problems and boundary spectral problem. Inverse Probl. **20**, 419–436 (2004)
45. Katsuda, A., Kurylev, Y., Lassas, M.: Stability of boundary distance representation and reconstruction of Riemannian manifolds. Inverse Probl. Imaging **1**, 135–157 (2007)
46. Krein, M.G.: Determination of the density of an inhomogeneous string from its spectrum (in Russian). Dokl. Akad. Nauk SSSR **76**(3), 345–348 (1951)
47. Kurylev, Y.: Multidimensional Gel'fand inverse problem and boundary distance map. In: Soga, H. (ed.) Inverse Problems Related to Geometry, pp. 1–15. Ibaraki University Press, Mito (1997)
48. Kurylev, Y., Lassas, M.: Hyperbolic inverse problem with data on a part of the boundary. In: Differential Equations and Mathematical Physics, Birmingham, 1999. AMS/IP Studies in Advanced Mathematics, vol. 16, pp. 259–272. AMS (2000)
49. Kurylev, Y., Lassas, M.: Hyperbolic inverse boundary-value problem and time-continuation of the non-stationary Dirichlet-to-Neumann map. Proc. R. Soc. Edinb. Sect. A **132**, 931–949 (2002)
50. Kurylev, Y., Lassas, M.: Inverse problems and index formulae for Dirac operators. Adv. Math. **221**, 170–216 (2009)
51. Kurylev, Y., Lassas, M., Somersalo, E.: Maxwell's equations with a polarization independent wave velocity: direct and inverse problems. J. Math. Pures Appl. **86**, 237–270 (2006)
52. Lasiecka, I., Triggiani, R.: Regularity theory of hyperbolic equations with nonhomogeneous Neumann boundary conditions. II. General boundary data. J. Differ. Equ. **94**, 112–164 (1991)
53. Lassas, M., Uhlmann, G.: On determining a Riemannian manifold from the Dirichlet-to-Neumann map. Ann. Sci. Ecole Norm. Super. **34**, 771–787 (2001)
54. Lassas, M., Sharafutdinov, V., Uhlmann, G.: Semiglobal boundary rigidity for Riemannian metrics. Math. Ann. **325**, 767–793 (2003)

55. Michel, R.: Sur la rigidité imposée par la longueur des géodésiques. Invent. Math. **65**, 71–83 (1981)
56. Mukhometov, R.G.: The reconstruction problem of a two-dimensional Riemannian metric, and integral geometry (Russian). Dokl. Akad. Nauk SSSR **232**(1), 32–35 (1977)
57. Mukhometov, R.G.: A problem of reconstructing a Riemannian metric. Sib. Math. J. **22**, 420–433 (1982)
58. Mukhometov, R.G., Romanov, V.G.: On the problem of finding an isotropic Riemannian metric in an n-dimensional space (Russian). Dokl. Akad. Nauk SSSR **243**(1), 41–44 (1978)
59. Otal, J.P.: Sur les longuer des géodésiques d'une métrique a courbure négative dans le disque. Comment. Math. Helv. **65**, 334–347 (1990)
60. Pestov, L., Uhlmann, G.: Two dimensional simple compact manifolds with boundary are boundary rigid. Ann. Math. **161**, 1089–1106 (2005)
61. Rachele, L.: An inverse problem in elastodynamics: determination of the wave speeds in the interior. J. Differ. Equ. **162**, 300–325 (2000)
62. Rachele, L.: Uniqueness of the density in an inverse problem for isotropic elastodynamics. Trans. Am. Math. Soc. **355**(12), 4781–4806 (2003)
63. Ralston, J.: Gaussian beams and propagation of singularities. In: Littman, W., Caffarelli, L.A. (eds.) Studies in Partial Differential Equations. MAA Studies in Mathematics, vol. 23, pp. 206–248. Mathematical Association of America, Washington, DC (1982)
64. Salo, M.: Stability for solutions of wave equations with C1,1 coefficients. Inverse Probl. Imaging **1**(3), 537–556 (2007)
65. Seeger, A., Sogge, C.D., Stein, E.M.: Regularity properties of Fourier integral operators. Ann. Math. **134**, 231–251 (1991)
66. Sharafutdinov, V.: Integral Geometry of Tensor Fields. VSP, Utrech (1994)
67. Smith, H.F.: A parametrix construction for wave equations with C1, 1 coefficients. Ann. Inst. Fourier Grenoble **48**(3), 797–835 (1998)
68. Smith, H.F.: Spectral cluster estimates for C1, 1 metrics. Am. J. Math. **128**(5), 1069–1103 (2006)
69. Smith, H.F., Sogge, C.D.: On the Lp norm of spectral clusters for compact manifolds with boundary. Acta Math. **198**, 107–153 (2007)
70. Stefanov, P., Uhlmann, G.: Rigidity for metrics with the same lengths of geodesics. Math. Res. Lett. **5**, 83–96 (1998)
71. Stefanov, P., Uhlmann, G.: Boundary rigidity and stability for generic simple metrics. J. Am. Math. Soc. **18**, 975–1003 (2005)
72. Stefanov, P., Uhlmann, G.: Local lens rigidity with incomplete data for a class of non-simple Riemannian manifolds. J. Differ. Geom. **82**, 383–409 (2009)
73. Stein, E.M.: Harmonic Analysis: Real-Variable Methods, Orthogonality, and Oscillatory Integrals. Princeton Mathematical Series, vol. 43. Monographs in Harmonic Analysis, III. Princeton University Press, Princeton (1993)
74. Sylvester, J.: An anisotropic inverse boundary value problem. Commun. Pure Appl. Math. **43**(2), 201–232 (1990)
75. Sylvester, J., Uhlmann, G.: A global uniqueness theorem for an inverse boundary value problem. Ann. Math. **125**, 153–169 (1987)
76. Sylvester, J., Uhlmann, G.: Inverse problems in anisotropic media. Contemp. Math. **122**, 105–117 (1991)
77. Tataru, D.: Unique continuation for solutions to PDEs, between Hörmander's theorem and Holmgren's theorem. Commun. Part. Differ. Equ. **20**, 855–884 (1995)
78. Tataru, D.: On the regularity of boundary traces for the wave equation. Ann. Scuola Norm. Sup. Pisa CL Sci. **26**, 185–206 (1998)
79. Tataru, D.: Unique continuation for operators with partially analytic coefficients. J. Math. Pures Appl. **78**, 505–521 (1999)
80. Tataru, D.: Strichartz estimates for operators with nonsmooth coefficients and the nonlinear wave equation. Am. J. Math. **122**(2), 349–376 (2000)

81. Tataru, D.: Strichartz estimates for second order hyperbolic operators with nonsmooth coefficients. II. Am. J. Math. **123**(3), 385–423 (2001)
82. Tataru, D.: Strichartz estimates for second order hyperbolic operators with nonsmooth coefficients. III. J. Am. Math. Soc. **15**, 419–442 (2002)
83. Uhlmann, G.: Developments in inverse problems since Calderón's foundational paper. In: Christ, M., Kenig, C., Sadosky, C. (eds.) Essays in Harmonic Analysis and Partial Differential Equations, Chap. 19. University of Chicago Press, Chicago (1999)
84. Wiechert, E., Zoeppritz, K.: Uber Erdbebenwellen. Nachr. Koenigl. Geselschaft Wiss. Goettingen **4**, 415–549 (2007)

Sonic Imaging

Frank Natterer

Contents

1 Introduction 1254
2 The Model Problem 1255
3 The Born Approximation 1255
An Explicit Formula for the Slab 1255
An Error Estimate for the Born Approximation 1258
4 The Nonlinear Problem in the Time Domain 1261
The Kaczmarz Method in the Time Domain 1261
Numerical Example (Transmission) 1263
Numerical Examples (Reflection) 1264
5 The Nonlinear Problem in the Frequency Domain 1265
Initial Value Techniques for the Helmholtz Equation 1265
The Kaczmarz Method in the Frequency Domain 1266
6 Initial Approximations 1266
7 Pecularities 1269
Missing Low Frequencies 1269
Caustics and Trapped Rays 1270
The Role of Reflectors 1271
8 Direct Methods 1273
Boundary Control 1273
Inverse Scattering 1275
9 Conclusion 1275
Cross-References 1275
References 1275

F. Natterer (✉)
Department of Mathematics and Computer Science, University of Münster, Münster, Germany
e-mail: nattere@math.uni-muenster.de

© Springer Science+Business Media New York 2015

O. Scherzer (ed.), *Handbook of Mathematical Methods in Imaging*,
DOI 10.1007/978-1-4939-0790-8_37

Abstract

This paper deals with the inverse problem of the wave equation, which is of relevance in fields such as ultrasound tomography, seismic imaging, and nondestructive testing. We study the linearized problem by Fourier analysis, and we describe an iterative reconstruction method for the fully nonlinear problem in the time domain. We discuss practical problems such as the spectral incompleteness in reflection imaging and finding a good initial approximation. We demonstrate by numerical reconstructions from synthetic data what can be achieved.

1 Introduction

Imaging with ultrasound, or sonography, is a well-established technique in medicine and many other fields. It goes back to the middle of the last century [21]. The human body is scanned by a (usually handheld) transducer that emits the sound signals and records the reflected signals. The mathematical models used in present days are fairly simple. In particular they assume straight line propagation of the sound signals. More refined models take into account the bending of the rays. They go under the name of diffraction tomography [15, 32, 56].

In this article we deal with an even more refined method of ultrasound imaging which is based on the wave equation. This permits to take into account not only bent rays but also multiple reflections. These techniques are not yet part of current medical practice, but prototype scanners based on these principles have already been built and tested in a clinical environment [4, 20].

Speaking in mathematical terms, we deal with the ultrasound imaging problem as the inverse problem for the wave equation. We determine parameters such as the speed of sound and the attenuation from measurements of the field outside the object. Imaging with the wave equation plays an important role not only in medical radiology, but also in nondestructive testing and seismic exploration. Our goal is to deal with the fully nonlinear problem. We start with a survey on the results obtained by linearization. For the fully nonlinear problem, we describe in detail the Kaczmarz method in the time domain. It turns out that Kaczmarz method, whose linear version is widely used in X-ray tomography [25], can be viewed in a very intuitive way as consecutive time reversal. We show by numerical examples that Kaczmarz method easily solves the standard problems, such as transmission tomography and reflection imaging with broadband data. We also study the behavior of the method in nonstandard situations, such as caustics, trapped rays, and, in particular, missing low frequencies in the source pulse.

The literature on inverse problems for the wave equation is inexhaustible. So we restrict our discussion to the exact solution of the fully nonlinear coefficient inverse problem for general objects. We do not even mention the important work on obstacle scattering [12], where the object is homogeneous and only the shape of the object is sought for, nor do we deal with level set methods [18]. Neither do we discuss the many approximate methods, except for the Born approximation, because it gives so

much insight into the fully nonlinear problem. Other problems which we ignore are the linear inverse source problems [29] and phase contrast tomography [31].

2 The Model Problem

We consider the following initial value problem for the wave equation. Let Ω be a domain in $R^n, n = 2, 3$ and let $\Gamma = \partial\Omega$. Let $c(x)$ be the speed of sound in $x \in \Omega$. Let $T > 0$ be the observation time. Let u be a solution of the initial-boundary value problem

$$\frac{\partial^2 u}{\partial t^2}(x,t) = c^2(x)\Delta u(x,t),\ x \in \Omega,\ 0 < t < T, \tag{1}$$

$$\frac{\partial u}{\partial \nu} = q(t)\delta_{\Gamma,s}(x),\ x \in \Gamma,\ 0 < t < T, \tag{2}$$

$$u(x,t) = 0,\ t < 0 \tag{3}$$

with ν the exterior normal on Γ. We consider the inverse problem of determining c from the values of $u(x,t)$ on $\Gamma \times (0,T)$, or on part of it. The function q represents the source pulse, and $\delta_{\Gamma,s}$ is the Dirac δ-function on Γ concentrated at the source $s \in \Gamma$. We always assume that $c^2(x) = c_0^2(x)/(1 + f(x))$ with the (known) background velocity c_0 and a function $f > -1$ which has to be determined.

We remark that everything we are doing in this paper can also be done for more general problems, such as problems with varying density [26], in moving media [61] and problems with attenuation [4, 20, 40].

3 The Born Approximation

Even though the underlying differential equation is linear, the inverse problem is nonlinear. Most of the literature on the problem makes use of some kind of linearization, mostly the Born approximation. Our goal is definitely to deal with the fully nonlinear problem. However linearization leads to valuable insights also for the nonlinear problem. Therefore we discuss the Born approximation in some detail.

An Explicit Formula for the Slab

We restrict ourselves to a very simple geometry: Ω is just the slab $0 < x_n < D$. Sources and receivers are sitting either on both boundaries (transmission mode) or on only one of them (reflection mode). It is more convenient to work with the inhomogeneous initial value problem

$$\frac{\partial^2 u}{\partial t^2}(x,t) = c^2(x)\left(\Delta u(x,t) + q(t)\delta(x-s)\right), \quad x \in R^n,\ 0 < t < T, \tag{4}$$

$$u(x,t) = 0, \quad t < 0 \tag{5}$$

with δ the Dirac function in R^n and s a source. We also assume $c(x) = c_0 > 0$ constant in this section. Equations (4), (5) is a simplified version of (1)–(3).

We also look at the problem in the frequency domain. For the Fourier transform in R^n, we use the notation

$$\hat{f}(\xi) = (2\pi)^{-n/2} \int_{R^n} e^{-ix\cdot\xi} f(x)dx$$

and we denote the inverse Fourier transform of f by $\tilde{f}$. Doing an inverse Fourier transform with respect to t in (4), we get

$$\Delta\tilde{u}(x,\omega) + k^2(1+f)\tilde{u}(x,\omega) = -\tilde{q}(\omega)\delta(x-s) \tag{6}$$

where $k = \omega/c_0$ is the wave number. Because of (5) this equation has to be complemented by the Sommerfeld radiation condition for outgoing radiation, i.e.,

$$\frac{\partial\tilde{u}}{\partial|x|} - ik\tilde{u} \to 0, |x| \to \infty.$$

In this section we derive an explicit representation of the Born approximation by Fourier analysis. We follow [14,42].

For $f = 0$, the solution of (6) is $\tilde{q}(\omega)G_k(x-s)$ with G_k the free space Green's function which we give in the plane wave decomposition

$$G_k(x) = ic_n \int_{R^{n-1}} e^{i(|x_n|\kappa(\xi')-x'\cdot\xi')} \frac{d\xi'}{\kappa(\xi')} \tag{7}$$

where $c_2 = 1/(4\pi)$, $c_3 = 1/(8\pi^2)$, $\kappa(z) = \sqrt{k^2 - z^2}$. For $z > k$, $\kappa(z)$ is defined to be $i\sqrt{z^2-k^2}$.

For arbitrary f, we put $\tilde{u}(x) = \tilde{q}(\omega)G_k(x-s) + v(x,s)$, obtaining for v

$$\Delta_x v(x,s) + k^2(1+f(x))v(x,s) = -k^2\tilde{q}G_k(x-s)f(x).$$

Neglecting fv leads to the Born approximation

$$\Delta_x v(x,s) + k^2 v(x,s) \approx -k^2\tilde{q}(\omega)G_k(x-s)f(x)$$

or

$$v(x,s) \approx k^2\tilde{q}(\omega)\int_\Omega G_k(x-y)G_k(y-s)f(y)dy.$$

With x' the first $n-1$ coordinates of x and correspondingly for s, y we rewrite this as

$$v(x', x_n, s', s_n) \approx k^2 \tilde{q}(\omega) \int_0^D \int_{R^{n-1}} G_k(x' - y', x_n - y_n) f(y', y_n) \\ G_k(y' - s', y_n - s_n) dy' dy_n.$$

Inserting (7) we get

$$v(x', x_n, s', s_n) \approx -c_n^2 k^2 \tilde{q}(\omega) \int_0^D \int_{R^{n-1}} \int_{R^{n-1}} e^{i(|x_n - y_n|\kappa(\xi') - (x'-y')\cdot\xi')} \\ \int_{R^{n-1}} f(y', y_n) e^{i(|y_n - s_n|\kappa(\sigma') - (y'-s')\cdot\sigma')} \frac{d\xi' d\sigma'}{\kappa(\xi')\kappa(\sigma')} dy' dy_n.$$

The y' integral is just a $(n-1)$-dimensional Fourier transform, hence

$$v(x', x_n, s', s_n) \approx -(2\pi)^{n-1} k^2 c_n^2 \tilde{q}(\omega) \int_0^D \int_{R^{n-1}} \int_{R^{n-1}} \\ e^{i(|x_n - y_n|\kappa(\xi') + |y_n - s_n|\kappa(\sigma') - x'\cdot\xi' + s'\cdot\sigma')} \hat{f}(\sigma' - \xi', y_n) \frac{d\xi' d\sigma'}{\kappa(\xi')\kappa(\sigma')} dy_n. \tag{8}$$

where $\hat{f}$ is the $(n-1)$-dimensional Fourier transform of f with respect to the first $n-1$ variables. Fourier transforms of dimensions $n-1$ with respect to x' and s' yield

$$\hat{v}(\xi', x_n, \sigma', s_n) \approx -k^2 c_n^2 \tilde{q}(\omega) \frac{(2\pi)^{2(n-1)}}{\kappa(\xi')\kappa(\sigma')} \int_0^D e^{i|x_n - y_n|\kappa(\xi') + i|y_n - s_n|\kappa(\sigma')} \\ \hat{f}(\xi' + \sigma', y_n) dy_n.$$

In particular we have

$$\hat{v}(\xi', 0, \sigma', 0) \approx A \hat{f}(\xi' + \sigma', -\kappa(\xi') - \kappa(\sigma')), \tag{9}$$

$$\hat{v}(\xi', D, \sigma', 0) \approx A e^{iD\kappa(\xi')} \hat{f}(\xi' + \sigma', \kappa(\xi') - \kappa(\sigma')) \tag{10}$$

where $\hat{f}$ is now the n-dimensional Fourier transform of f and $A = -k^2 \tilde{q}(\omega) c_n^2 (2\pi)^{3n/2-1} / (\kappa(\xi')\kappa(\sigma'))$. Equations (9), (10) is the solution of the inverse problem of reflection and transmission imaging, resp., in the Born approximation. Note that the derivation of (9), (10) depends entirely on the plane wave decomposition of G_k which was used first in this context in [14].

Let's look at the transmission case, i.e., (10). For simplicity, we consider the 2D case. Consider the semicircle of radius k in the upper half plane around the origin. Attach to each point of this semicircle the semicircle of radius k around this point that opens downward. According to (10), $\hat{f}$ is determined by the data along each

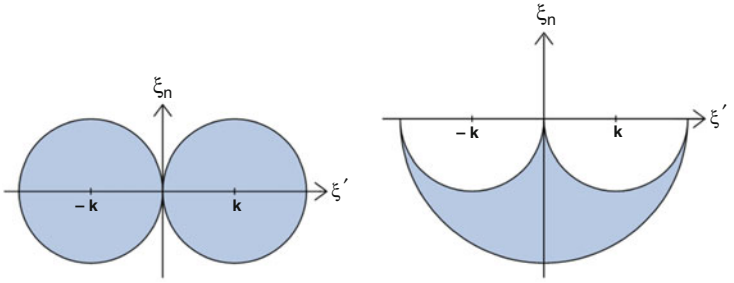


Fig. 1 *Left*: Fourier coverage for transmission imaging. *Right*: Fourier coverage for reflection imaging. k is the wave number. The figures are for $n = 2$. For $n = 3$, one has to rotate the figures around the vertical axis

of these semicircles. These semicircles fill the circles of radius k with midpoints $(\pm k, 0)$; see Fig. 1. In the reflection case (9) we proceed in the same way, except that we start out from the semicircle in the lower half plane. Now the attached semicircles fill the semicircle of radius $2k$ in the lower half plane, except for the circles of radius k around $(\pm k, 0)$.

Thus we see a fundamental difference between transmission and reflection imaging: In transmission imaging low frequency features f can be recovered irrespectively of the frequency content of the source pulse q, whereas in reflection imaging one needs low frequencies in the source pulse to recover low frequency features in f. This is one of the main difficulties in seismic imaging; see [9, 19, 22, 30,60]. To the best of our knowledge, Fig. 1 appeared first in [35,64]. The difficulties in reflection imaging without low frequencies will be discussed in section "Missing Low Frequencies".

We also conclude from (9), (10) that combined transmission and reflection measurements determine $\hat{f}$ within the ball of radius $2k$. In other words, since $\lambda = 2\pi/k$ is the wavelength of the irradiating sources, the spatial resolution according to Shannon's sampling theorem is $2\pi/2k = \lambda/2$. This is a classical result already known to Born.

The numerical evaluation of (9), (10) is by no means trivial. In principle it can be done by Fourier transforms. However this requires a sophisticated software for non-equispaced fast Fourier transforms. Assuming uniform sampling of $\hat{v}$ in ξ' and σ' leads to non-equispaced sampling of $\hat{f}$ on a grid of the type shown in Fig. 2. An implementation similar to the filtered backprojection algorithm of X-ray computerized tomography has been given in [14].

An Error Estimate for the Born Approximation

We now assume that Ω is a ball of radius r, the background speed is constant as in the previous section, and that the illumination done is by plane waves. This leads to

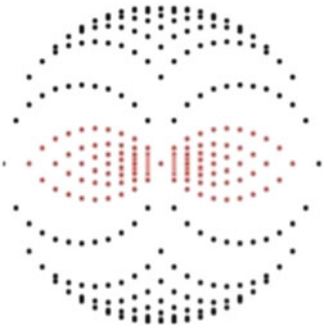


Fig. 2 Non-equidistant grid at which $\hat{f}$ is given by (9), (10) for $n = 2$

the problem

$$\Delta u + k^2(1 + f)u = 0, \tag{11}$$

$$u = u_i + u_s \tag{12}$$

where $u_i = \exp ikx \cdot \theta$ is the incoming wave with direction $\theta \in S^{n-1}$ and wave number k and u_s is the scattered wave satisfying the Sommerfeld radiation condition of exterior radiation. We assume f to be supported in Ω. θ runs through all of S^1 for $n = 2$ and through a great circle of S^2 for $n = 3$. The problem is to determine f from $g_\theta(z) = u_s(s\theta + z)$ for some s with $|s| > r$ and $z \in \theta^\perp$. The decisive tool for the error estimate is an estimate for the operator

$$\mathcal{G}_k u(x) = k^2 \int G_k(x - y)u(y)dy$$

in $L_2(|x| < r)$. In [41] it has been shown that

$$||\mathcal{G}_k||_{L_2(|x|<r)} = kr\gamma_n(kr)$$

where γ_n is a function with $\bar{\gamma}_n = \sup_{(0,\infty)} \gamma_n < \infty$. For instance, $\bar{\gamma}_2 \leq 0.8$. We have $u_s = k^2\mathcal{G}_k(f(u_i + u_s))$, hence, with $q = kr\gamma_n(kr)$, $q||f||_{L_\infty(|x|<r)} < 1$

$$||u_s||_{L_2(|x|<r)} \leq \frac{q}{1 - q||f||_{L_\infty(|x|<r)}} ||f||_{L_2(|x|<r)}. \tag{13}$$

We introduce the propagation operator

$$(U_\theta f)(z) = \int G_k(s\theta + z - y) f(y)u_i(y)dy;$$

see [14]. For the $(n-1)$-dimensional Fourier transform of U_θ in $\theta^\perp$, we have

$$(U_\theta f)^\wedge(\zeta) = i\sqrt{\frac{\pi}{2}}e^{i|s|\kappa(\zeta)}\frac{1}{\kappa(\zeta)}\hat{f}((\epsilon\kappa(\zeta)-k)\theta+\zeta)$$

where $\kappa(\zeta) = \sqrt{k^2-|\zeta|^2}$ and $\epsilon = \operatorname{sgn}(s)$. This is essentially Theorem 3.1 of [50].

We have

$$k^{-2}g_\theta = U_\theta f + U_\theta\frac{u_s}{u_i}f$$

hence

$$k^{-2}\hat{g}_\theta(\zeta) = i\sqrt{\frac{\pi}{2}}e^{i|s|\kappa(\zeta)}\frac{1}{\kappa(\zeta)}\left(\hat{f}((\epsilon\kappa(\zeta)-k)\theta+\zeta) + (u_s f)^\wedge((\epsilon\kappa(\zeta)\theta+\zeta)\right).$$

Now let $\xi \in R^n$ be in the ball $|\xi| < 2k$. We can find $\theta \in S^{n-1}$, $\zeta \in \theta^\perp$ with $\xi = (\epsilon\kappa(\zeta)-k)\theta+\zeta$. For any such choice of θ and ζ, the last formula yields

$$\hat{f}(\xi) = \hat{f}_B(\xi) - (u_s f)^\wedge(\xi+k\theta), \quad \hat{f}_B(\xi) = -i\sqrt{\frac{2}{\pi}}e^{-i|s|\kappa(\zeta)}\kappa(\zeta)k^{-2}\hat{g}_\theta(\zeta).$$

Since

$$|(u_s f)^\wedge| \le (2\pi)^{-n/2}||u_s||_{L_2(|x|<r)}||f||_{L_2(|x|<r)}$$

we have for $|\xi| \le 2k$

$$\left|\hat{f}(\xi) - \hat{f}_B(\xi)\right| \le (2\pi)^{-n/2}\frac{q}{1-q||f||_{L_\infty(|x|<r)}}||f||^2_{L_2(|x|<r)}. \tag{14}$$

Thus we obtained an approximation of second order in f for $\hat{f}$ in the ball $|\xi| \le 2k$. This is what we expect from the Born approximation. The restriction to the bandwidth $2k$ is also natural as this corresponds to the maximal spatial resolution of π/k.

We see from (14) that the decisive quantity for the Born approximation to be valid is $q||f||_{L_\infty(|x|<r)}$. In the engineering literature, see, e.g., [32], it is well known that this is a measure for the phase shift generated by the object f. Thus we arrive at the conclusion that the Born approximation works if the phase shift is sufficiently small. We also conclude from (14) that the reconstruction process, if restricted to the resolution π/k is fairly stable: there are no unstable operations in the computation of $\hat{f}_B$. It has been shown in [53] that this is also the case for the fully nonlinear problem, provided the geodesics behave reasonably. Thus the frequently discussed

instability of the inverse scattering problem is just a consequence of an inadequate choice of the frequency of the irradiating waves.

4 The Nonlinear Problem in the Time Domain

In this section we deal with the problem in the form (1)–(3). We suggest a very simple but practically useful iterative method and discuss some theoretical and practical aspects of it.

The Kaczmarz Method in the Time Domain

For a finite number of sources s_j, $j = 1, \ldots, p$ we have to solve a system of the form

$$R_j(f) = g_j, \ j = 1, \ldots, p. \tag{15}$$

where R_j is an operator between a Hilbert space H of functions on $\Omega \times (0, T)$ and certain Hilbert spaces H_j of functions on $\Gamma \times (0, T)$. In our case the operator R_j is defined by $R_j(f) = u_{j\,|\Gamma\times(0,T)}$, where u_j is the solution of (1)–(3) for the source $s = s_j$.

A natural method for solving (15) is the Kaczmarz method. This is an iterative method with the update

$$f_{j+1} = f_j + \alpha(R'_{j'}(f_j))^* C_{j'}^{-1}(g_{j'} - R_{j'}(f_j)), \quad j = 1, 2, \ldots$$

where $j' = j \bmod p$. Here, R'_j is the Fréchet derivative of R_j, ${R'_j}^*$ its adjoint, and C_j is a positive definite operator. Going through all the equations once is called a complete sweep of the Kaczmarz method.

In the linear case, i.e., if $R_j : H \to H_j$ is a linear bounded operator the convergence of the Kaczmarz method is well understood, in particular for the case of CT; see, e.g., [50], Theorem 5.1.

If $C_j - R_j R_j^*$ is positive semidefinite, the sequence f_j converges for $0 < \alpha < 2$ to $Pf_1 + R^+ g$, where P is the orthogonal projection on the nullspace of $R = (R_1, \ldots, R_p)$ and R^+ is the Moore–Penrose generalized inverse of R. In the nonlinear case the situation is much less clear, see [6] for a theoretical discussion of the issue.

The Kaczmarz method is not to be confused with the Landweber iteration. In the Landweber method one does the update only after all the equations have been processed, while in Kaczmarz we do the update as soon as a single equation is processed. In practice Kaczmarz turns out to be much faster than Landweber. In the linear finite dimensional case Kaczmarz is identical to the SOR method, see Theorem 5.2 in [50]. Also, the speed of convergence depends on the ordering

of the sources. Sequential ordering does not yield the best results. Sophisticated arrangements, in particular random orderings, are much better. For the case of CT, see Section 5.3.1 of [50].

In order to compute $R_j'(f)$ for our imaging problem, we replace f by $f+h$ and u_j by u_j+w with h, w small and ignore higher order terms. We get for w the differential equation

$$\frac{1}{c^2}\frac{\partial^2 w}{\partial t^2} = \Delta w - \frac{h}{c_0^2}\frac{\partial^2 u_j}{\partial t^2}, \quad x\in\Omega,\ 0<t<T, \tag{16}$$

$$\frac{\partial w}{\partial \nu} = 0, \quad x\in\Gamma,\ 0<t<T, \tag{17}$$

$$w = 0, \quad t<0. \tag{18}$$

Hence we expect that

$$R_j'(f)h = w_{|\Gamma\times(0,T)}.$$

In fact it has been shown in [16] that with a suitable choice of Hilbert spaces H, H_j the operator R_j is Fréchet differentiable and the expression above is the derivative. A possible choice is $H = H^2(\Omega)$, restricted to functions $f > -1$, and $H_j = H^{1/2}(\Gamma\times(0,T))$.

Now we come to the problem of computing the adjoint of R_j'. For this we have to specify the spaces H, H_j. To make things easy, we put $H = L_2(\Omega), H_j = L_2(\Gamma\times(0,T))$, i.e., we consider R_j' as an unbounded operator between these spaces. Then the adjoint ${R_j'}^*$ satisfies

$$(R_j'(f)h, g)_{L_2(\Omega\times(0,T))} = (h, R_j'(f)^* g)_{L_2(\Gamma\times(0,T))}$$

for sufficiently smooth functions h, g. Determining the exact domain of definition of the adjoint is a more tricky question which is not considered here.

To find an explicit form of the adjoint, we follow [38]. We start out from the identity

$$\begin{aligned}\int_\Omega\int_0^T \left(\frac{1}{c^2}\frac{\partial^2 w}{\partial t^2} - \Delta w\right) z\,dt\,dx - \int_\Omega\int_0^T \left(\frac{1}{c^2}\frac{\partial^2 z}{\partial t^2} - \Delta z\right) w\,dt\,dx \\ = \int_\Gamma\int_0^T \left(\frac{\partial z}{\partial \nu}w - z\frac{\partial w}{\partial \nu}\right) dt\,dx + \left[\int_\Omega \frac{1}{c^2}\left(\frac{\partial w}{\partial t}z - w\frac{\partial z}{\partial t}\right) dx\right]_0^T.\end{aligned}$$

where ν is the exterior normal on Γ. This holds for any sufficiently smooth functions w, z on $\Omega\times(0,T)$. Choosing for w the solution of (16)–(18) with $u = u_j$ the solution of (1)–(3) for source j and for z the solution of the final value problem

$$\frac{\partial^2 z}{\partial t^2} = c^2 \Delta z \text{ in } \Omega \times (0,T), \tag{19}$$

$$\frac{\partial z}{\partial \nu} = g \text{ on } \Gamma \times (0,T), \tag{20}$$

$$z = 0 \text{ for } t > T \tag{21}$$

we obtain

$$-\int_\Omega \frac{h}{c_0^2} \int_0^T \frac{\partial^2 u_j}{\partial t^2} z dt dx = \int_\Gamma \int_0^T g(R_j'(f)h) dt dx,$$

or, using inner products,

$$-\left(h, \frac{1}{c_0^2}\int_0^T \frac{\partial^2 u_j}{\partial t^2} z dt\right)_{L_2(\Omega)} = \left(R_j'(f)h, g\right)_{L_2(\Gamma\times(0,T))}.$$

Hence

$$(R_j'(f))^* g = -\frac{1}{c_0^2}\int_0^T \frac{\partial^2 u_j}{\partial t^2} z dt \tag{22}$$

with z the solution of the final value problem (19)–(21). Note that z is nothing but the time reversed field. Thus the Kaczmarz method is just a consecutive form of time reversal, a technique that is used extensively for imaging problems [2, 3, 57].

Numerical Example (Transmission)

In this section we show what can be achieved for a mammography scanner patterned after [20] for the Salt Lake City breast phantom suggested in [4]. The reconstruction region is a circle of radius 8 cm. On the boundary we have 128 receivers and a modest number of sources, namely 8, see Fig. 3. The source pulse at the sources is $q(t) = \cos(\omega t)\exp(-(t/\tau)^2)$ with $\tau = \pi/\omega$. The frequency of the irradiating waves is 1 MHz, i.e., $\omega = 2\pi 10^6$/s. With a background speed $c_0 = 1{,}500$ m/s this corresponds to a wavelength of 1.5 mm, hence to a spatial resolution of 0.75 mm. In Fig. 4 we display the rays, suggesting that a straight line assumption is useless. We show the reconstruction after 1 sweep and after 6 sweeps of the Kaczmarz method with relaxation parameter $\alpha = 2 \times 10^{12}$ and $C_j = 1$. Note that the number of sources as well as the number of sweeps is very small, leading to very reasonable reconstruction times. The spatial stepsize was chosen to be 0.33 mm, which corresponds to 1/6 of the wavelength.

For truly 3D examples, see [23, 39].

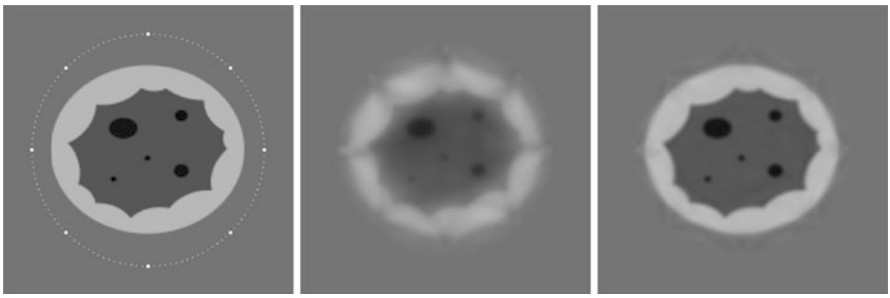


Fig. 3 Kaczmarz method in time domain for breast phantom. *Left*: Phantom with 8 sources and 128 receivers. *Middle*: One sweep of the Kaczmarz method. *Right*: Six sweeps of the Kaczmarz method. *Grey* window in all panels is from -0.1 to 0.1

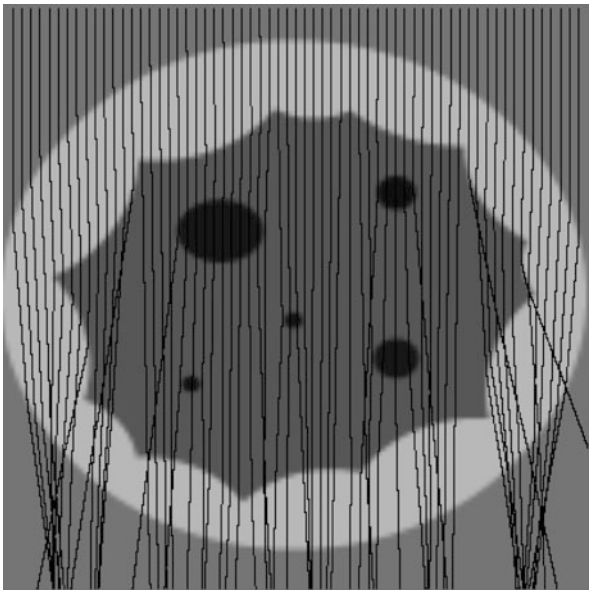


Fig. 4 Rays for breast phantom

Numerical Examples (Reflection)

In this example the reconstruction region is the slab $0 \leq x_n \leq D$, and we have sources and receivers on $x_n = 0$ only. This is the situation, e.g., in seismic imaging and in supine mammography [17, 26, 44]. In our example we adjust the Salt Lake City breast phantom of the previous section to the supine position of the patient, see Fig. 5. The five tumors have a diameter of 1.5 mm. To discover them, we need incoming waves of wavelength 3 mm, which, assuming a background speed of sound of 1,500 m/s, corresponds to a frequency of 500 kHz. From the section on the Born approximation we know that for reflection imaging we need low frequencies. So we are not surprised that the Kaczmarz method, as probably any other iterative

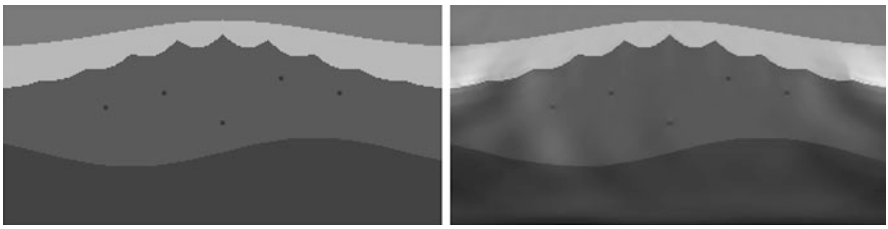


Fig. 5 *Left*: Breast phantom for patient in supine position. *Right*: Reconstruction

method, does not work for the comparatively high frequency of 500 kHz. What we do to circumvent this difficulty is to start the iteration with data for 30 kHz, use the result as an initial approximation for 1.2 MHz, and use the result again as an initial approximation for the final reconstruction with 500 kHz. This idea of starting with low frequencies is due to [8]. On each level we do only 6 sweeps, with relaxation parameters α chosen by trial and error. From Fig. 5 we see that the tumors are easily recovered, as we expect from the theory. Height of the pictures is 7.5 cm. The numerical work has been done on a 20 cm $\times$ 10 cm rectangle with a step size of 0.5 mm. We remark that only 19 sources sitting on the top boundary of the reconstruction region were used, but the number of receivers (400) was much larger. In the computation we did not use these 19 sources consecutively. We rather chose the sources in a random order.

5 The Nonlinear Problem in the Frequency Domain

This problem was considered in the early nineties as the paradigm of a nonlinear ill-posed problem. It was treated by regularized versions of Newton's method. Numerical examples were restricted to objects of the size of the wavelength, see [24, 27, 33]. However in medical applications, e.g., in mammography, the object size is typically 100 wavelengths.

Initial Value Techniques for the Helmholtz Equation

In order to carry out the Kaczmarz method in the frequency domain, we have to solve the forward problem (6) and the corresponding adjoint problem repeatedly. In particular for high wave numbers k this is a numerical challenge [27]. Fortunately it is possible to use initial value techniques for the solution of elliptic equations such as (6). Such methods are notoriously unstable. However it is possible to stabilize them, simply by removing the evanescent waves, see [34, 48, 51]. These methods are effective only if the true speed of sound deviates from the background only a little, typically by not more than 10 %, as is the case in mammography. An improved version that is able to handle much greater variations of the velocity, as they occur, e.g., in seismics, can be found in [59].

The Kaczmarz Method in the Frequency Domain

Based on the initial value techniques of the previous section, an extremely efficient implementation of the Kaczmarz method in the frequency domain is possible [49]. We describe the method for Ω the ball of radius r and plane wave irradiation. f may now be a complex-valued function

$$f(x) = \frac{c_0^2}{c(x)^2} - 1 - \frac{i}{k}\frac{2\gamma(x)c_0}{c(x)}$$

with $\gamma(x)$ the (unknown) attenuation. Let $\Gamma_j^{\pm}$ be the planes or straight lines perpendicular to θ_j touching Γ. We assume the functions $g_j^{\pm} = u_s$ on $\Gamma_j^{\pm}$ to be known, i.e., $\Gamma_j^{\pm}$ play the role of the detectors. The operator $R_j : L_2(\Omega) \to L_2(\Gamma_j^+)$, where u_s is the solution of the initial value problems for (11) with

$$u_s = g_j^-$$
$$\frac{\partial \tilde{u}_s}{\partial x_n}(x') = (2\pi)^{(1-n)/2} \int_{R^{(n-1)}} i\kappa(\xi')\hat{\tilde{u}}_s(\xi')e^{-ix'\cdot\xi'}d\xi'$$

on Γ_j^-, where x', ξ' are the local variables on Γ_j^-, $\kappa(\xi') = \sqrt{k^2 - |\xi'|^2}$ and $\hat{}$ stands for the $(n-1)$-dimensional Fourier transform with respect to x'. The last equation is a finite form of the Sommerfeld radiation condition, see, e.g., [11]. As in section (4.1), see also [38], we compute the adjoint of the derivative $(R_j'(f))^* : L_2(\Gamma_j^+) \to L_2(\Omega)$: $(R_j'(f)) * (r) = z$, where z is the solution of

$$\Delta z + k^2(1+f)z = 0 \text{ between } \Gamma_j^- \text{ and } \Gamma_j^+, \tag{23}$$

$$z = 0, \quad \frac{\partial z}{\partial \theta_j} = \overline{ru_s} \text{ on } \Gamma_j^+. \tag{24}$$

We remark that (23), (24) is just the frequency domain form of time reversal.

6 Initial Approximations

The biggest problem with Kaczmarz method is to find an initial approximation for which the process converges. In this section we derive a heuristic criterion that works surprisingly well in practice, at least for transmission imaging. The condition reads

$$\left| \int (f - f_0) ds \right| \le \lambda \tag{25}$$

where λ is the largest wavelength contained in the spectrum of the pulse q and the integral is taken along the geodesics belonging to the initial guess c_0 for c.

We give two derivations of (25). Both of them are heuristic but fortunately they agree, and they perform well in practice. The first derivation uses the frequency domain formulation (6) of the problem. The forward operator R_j is defined by $R_j(f) = \tilde{u}|_\Gamma$ with $\tilde{u}$ the solution of (6) with $s = s_j$. With $w = \tilde{u} - \tilde{u}_0$, $\tilde{u}_0$ the solution of (6) with $f = f_0,\ s = s_j$ we have

$$R_j(f) - R_j(f_0) = w|_\Gamma, \tag{26}$$

$$-\Delta w - k^2(1 + f_0)w = k^2(f - f_0)\tilde{u}. \tag{27}$$

For simplicity, we assume that we do the Kaczmarz method with $C_j = R_j'(f_j)(R_j'(f_j))^*$. Then the first approximation f_1 of the Kaczmarz method satisfies

$$R_j'(f_0)(f_1 - f_0) = \alpha(R_j(f) - R_j(f_0)).$$

Let w_1 be the solution of

$$-\Delta w_1 - k^2(1 + f_0)w_1 = k^2(f_1 - f_0)\tilde{u}_0/\alpha. \tag{28}$$

Then

$$w_1|_\Gamma = R_j'(f_0)(f_1 - f_0)/\alpha = R_j(f) - R_j(f_0).$$

Thus the solutions $w,\ w_1$ of (27), (28) coincide on Γ. Comparing (27) and (28), we see that f_1 can be similar to f only if $\tilde{u},\ \tilde{u}_0$ have similar phase,

$$|\text{phase}(\tilde{u}) - \text{phase}(\tilde{u}_0)| < \pi, \tag{29}$$

say. If this condition is not satisfied, then even the first iterate f_1 goes astray. Thus we consider (25) as a highly necessary condition for the convergence of the Kaczmarz method with starting element f_0. According to the WKB-approximation of $\tilde{u}$, we have

$$\tilde{u} \approx A \exp(ik\Phi)$$

where the phase Φ satisfies the eikonal equation $|\nabla\Phi|^2 = 1 + f$ and A is the real-valued amplitude. The same holds for $\tilde{u}_0$ with an amplitude A_0 and a phase Φ_0. For f close to f_0, we have

$$\Phi \approx \Phi_0 + \frac{1}{2}\int (f - f_0)ds$$

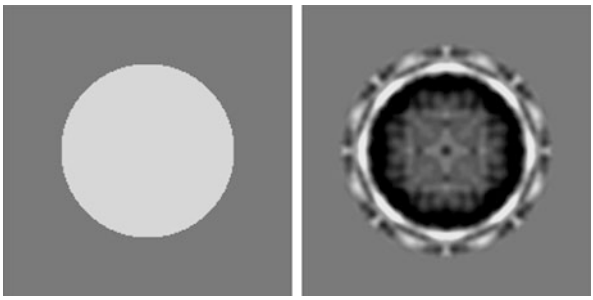


Fig. 6 Example for non-convergence if (25) is not satisfied. *Left*: True object. *Right*: Object to which the method converges

where the integral is along the geodesics of the background; see [53]. Combining the last three equations leads to (25).

The second derivation of (25) takes place in the time domain. We look at the traces $u(x', 0, t)$ and $u_0(x', 0, t)$. Qualitatively they look similar: They resemble time delayed and scaled versions of the pulse q. The delays are roughly the integrals of $1/c, 1/c_0$, resp., along the geodesics. Thus the difference of the delays is approximately

$$\int \frac{ds}{c} - \int \frac{ds}{c_0} = \int \left(\sqrt{1+f} - \sqrt{1+f_0}\right) \frac{ds}{c_0} = \int \frac{f - f_0}{2} \frac{ds}{c_0}.$$

The width w of the pulse q is determined by the lowest frequency ω in the pulse. We have approximately $w = \pi/\omega$. Minimization of the norm of the difference of the traces has an effect only if the traces overlap, i.e., if the difference of the delays is smaller that the width of the pulse. This means that

$$\left| \int \frac{f - f_0}{2} \frac{ds}{c_0} \right| < \pi/\omega$$

Since $\lambda = 2\pi c_0/\omega$, this is equivalent to (25).

We remark that (25) has been derived in [54] as the condition for the validity of the ray method for the inverse problem.

In order to confirm (25) numerically we created a simple test phantom which is 0.125 in the circle of radius 6 cm, the ambient speed of sound c_0 being 1,500 m/s. It was illuminated by 8 sources sitting on the circle of radius 8 cm. The sources had a constant signature from 350 down to 120 and 80 kHz, respectively. In this case the left hand side of (25) is 1.5 cm for $f_0 = 0$, while the largest wavelength is 1.25 and 1.875 cm, respectively. Thus (25) is satisfied for the lowest frequency 80 kHz, and in fact we observed convergence to the true object. For the lowest frequency 120 kHz (25) is not satisfied. The method became stationary at an object that is quite different from the true object; see Fig. 6. We remark the highest frequency (350 kHz) is of no relevance here. If we double it, the convergence behavior is the same.

7 Pecularities

Missing Low Frequencies

Often the source pulse q contains only high frequencies, i.e., $\tilde{q}$ (the signature in the jargon of seismic imaging) vanishes near 0. As we have seen previously this is not much of a problem in transmission imaging; it only forces us to use a good initial approximation. But it is one of the main difficulties in reflection imaging. We mentioned this in the framework of the Born approximation in Sect. 3. Again we restrict ourselves to the two-dimensional case. Assuming that the source pulse q has wave numbers in $[k_{\min}, k_{\max}]$, we conclude from (9) that $\hat{f}$ is determined by the reflection data inside the circle of radius $2k_{\max}$, except for the circles with radius $k_{\min}$ around $(\pm k_{\min}, 0)$; see Fig. 7. This means that the reconstruction is essentially a high pass filtered version of the true object.

For special media, we can do better [47]. For instance, if the medium in the slab $0 < x_n < D$ is layered, i.e., f depends on x_n only, then the n-dimensional Fourier transform of f becomes $\hat{f}(\xi', \xi_n) = (2\pi)^{(n-1)/2}\delta(\xi')\hat{f}(\xi_n)$, where δ is the Dirac δ function in R^{n-1} and $\hat{f}$ on the right hand side is the one-dimensional Fourier transform. In the reflection case we have $x_n = s_n = 0$ in (8), and since the problem is invariant with respect to translation in R^{n-1} we can restrict s' to 0. Hence (8) assumes the form

$$v(x', 0, 0, 0) \approx -c_n^2 k^2 (2\pi)^{n-1/2} \tilde{q}(\omega) \int_{R^{n-1}} e^{-ix'\cdot\xi'} \hat{f}(-2\kappa(\xi')) \frac{d\xi'}{k^2 - |\xi'|^2}$$

or

$$(\Delta_{x'} + k^2) v(x', 0, 0, 0) \approx c_n^2 k^2 (2\pi)^{n-1/2} \tilde{q}(\omega) \int_{R^{n-1}} e^{-ix'\cdot\xi'} \hat{f}(-2\kappa(\xi')) d\xi'. \tag{30}$$

If x' runs through all of R^{n-1}, we can recover $\hat{f}$ from this relation in the interval $[0, 2k]$. However in practice x' is restricted to a finite aperture $|x'| \le R$. This makes

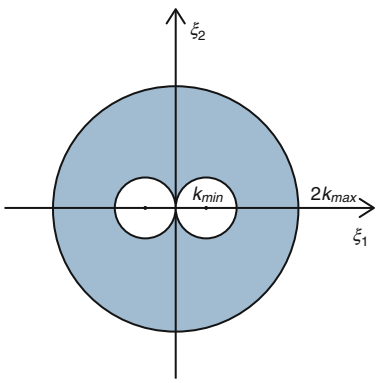


Fig. 7 Fourier coverage for reflection data with smallest wavenumber $k_{\min}$ and largest wavenumber $k_{\max}$

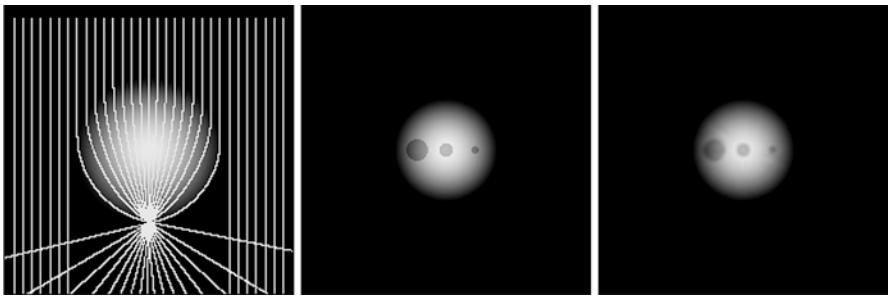


Fig. 8 *Left*: Luneburg lens with rays focussing on a focal point on the rim. *Middle*: Luneburg lens with a few tumors. *Right*: Reconstruction with 6 sweeps of the Kaczmarz method

it difficult to determine $\hat{f}$ in all of $[0, 2k]$. For instance, let f represent a point object in depth z, i.e., $f(x_n) = \delta(x_n - z)$. Then, $\hat{f}(-2\kappa(\xi')) = (2\pi)^{-1/2} e^{2i\kappa(\xi')}$ is a smooth function away from $|\xi'| = k$, but oscillates rapidly near this manifold, in particular for large z. In the vicinity of a point ξ', $|\xi'| < k$ it looks like a function of bandwidth

$$-2z|\nabla_{\xi'}\kappa(\xi')| = 2z|\xi'|/\kappa(\xi').$$

Thus, by the sampling theorem, $\hat{f}(\xi')$ can be recovered from (30) only if $R > 2z|\xi'|/\kappa(\xi')$, i.e., for $|\xi'| < k/\sqrt{1 + 4z^2/R^2}$. Summing up we arrive at the following conclusion:

We can determine $\hat{f}(\xi')$ for objects up to depth z from reflection data with aperture R and wave number k only for

$$k/\sqrt{1 + 4z^2/R^2} \le |\xi'| \le 2k. \tag{31}$$

The derivation for (31) given here is entirely heuristic, but it has been corroborated in many numerical experiments. For an alternative derivation, see [62]. Making use of analytic continuation one can go beyond this result and determine the remaining part of $\hat{f}$ provided that k, R, z satisfy a certain condition, see [47].

Another case in which we can do something without low frequencies is the case of media with a small dip angle, see [46].

Caustics and Trapped Rays

It seems that the Kaczmarz method is very robust with respect to pecularities, such as caustics and trapped rays. As an example in which caustics occur we treat the famous Luneburg lens [36], which generates a focal spot on the rim of the lens, see Fig. 8. We see that the Kaczmarz method does not have the slightest problem with the reconstruction.

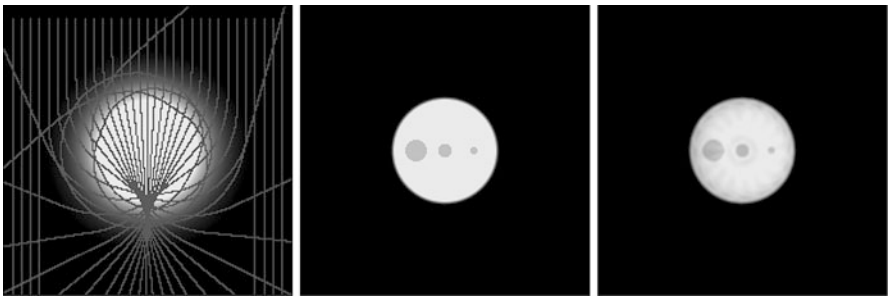


Fig. 9 *Top left*: Medium with trapped rays. *Middle*: Trapped rays medium with tumors. *Right*: Reconstruction with 6 sweeps of the Kaczmarz method

With trapped rays we have the same situation. In Fig. 9 we display the ray pattern for a medium with trapped rays [28]. We add some tumors and do the reconstruction. As in the previous example the Kaczmarz method does the reconstruction without any trouble.

The Role of Reflectors

It belongs to the lore of seismic imaging that reflectors greatly improve the reconstructed images; see [10, 13, 63]. In this section we study this phenomenon within the Born approximation and we demonstrate its practical relevance by numerical simulations; see also [45].

We again assume Ω to be the strip $0 < x_n < D$. We use the form (4), (5) of the inverse problem. We put the sources on $x_n = D$ and model the reflector at $x_n = 0$ by stipulating the Neumann boundary condition $\partial u/\partial x_n = 0$ on $x_n = 0$. Proceeding exactly as in section "An Explicit Formula for the Slab" we get for the Born approximation

$$v(x', x_n, s', s_n) \approx k^2 \tilde{q}(\omega) \int_\Omega G_k^0(x' - y', x_n, y_n) G_k^0(y' - s', y_n, s_n) f(y) dy.$$

where

$$G_k^0(x' - y', x_n, y_n) = G_k(x' - y', x_n - y_n) + G_k(x' - y', x_n + y_n)$$

is Green's function for the Neumann boundary condition at $x_n = 0$. We first do $(n-1)$-dimensional Fourier transforms with respect to x' and s', obtaining

$$\hat{v}(\xi', x_n, \sigma', s_n) \approx k^2 \tilde{q}(\omega) \int_0^D \int_{R^{n-1}} \hat{G}_k^0(\xi', x_n, y_n) \hat{G}_k^0(\sigma', y_n, s_n) e^{-i(\xi' + \sigma') \cdot y'} dy' dy_n \tag{32}$$

where the hat stands for the $(n-1)$-dimensional Fourier transform. From (7) we see that

$$\hat{G}_k(\xi', x_n) = i(2\pi)^{(n-1)/2} c_n \frac{e^{i|x_n| a(\xi')}}{\kappa(\xi')}.$$

Thus,

$$\hat{G}_k^0(\xi', x_n, y_n) = i(2\pi)^{(n-1)/2} c_n \frac{e^{i|x_n - y_n|\kappa(\xi')} + e^{i|x_n + y_n|\kappa(\xi')}}{\kappa(\xi')}.$$

It follows that for $x_n = s_n = D,\ 0 \le y_n \le D$

$$\hat{G}_0(\xi', x_n, y_n) = 2i(2\pi)^{(n-1)/2} c_n \frac{e^{i r_n \kappa(\xi')}}{\kappa(\xi')} \cos(y_n \kappa(\xi'))$$

$$\hat{G}_0(\sigma', y_n, s_n) = 2i(2\pi)^{(n-1)/2} c_n \frac{e^{i s_n \kappa(\sigma')}}{\kappa(\sigma')} \cos(y_n \kappa(\sigma'))$$

Inserting this into (32) leads to

$$\hat{v}(\xi', D, \sigma', D) \approx A \int_0^D \int_{Rn-1} e^{-i(\xi' + \sigma')\cdot y'} \cos(\kappa(\xi') y_n) \cos(\kappa(\sigma') y_n) f(y', y_n)$$
$$dy' dy_n,$$
$$A = -4c_n^2 (2\pi)^{n-1} k^2 \tilde{q}(\omega) \frac{e^{i(D\kappa(\xi') + D\kappa(\sigma'))}}{\kappa(\xi')\kappa(\sigma')}.$$

The y' integral is a $(n-1)$-dimensional Fourier transform. Hence

$$\hat{v}(\xi', D, \sigma', D) \approx (2\pi)^{(n-1)/2} A \int_0^D \cos(\kappa(\xi') y_n) \cos(\kappa(\sigma') y_n) \hat{f}(\xi' + \sigma', y_n) dy_n. \tag{33}$$

Here $\hat{f}$ is the $(n-1)$-dimensional Fourier transform of f with respect to x'. Since

$$\cos(\kappa(\xi') y_n) \cos((\kappa(\sigma') y_n) = \tfrac{1}{2}\{\cos((\kappa(\xi') + \kappa(\sigma')) y_n) + \cos((\kappa(\xi') - \kappa(\sigma')) y_n)\}$$

we have

$$\hat{v}(\rho', D, \sigma', D) \approx (2\pi)^{(n-1)/2} \frac{A}{2} \{(C\hat{f}(\xi' + \sigma', \cdot))(\kappa(\xi') + \kappa(\sigma')) + (C\hat{f}(\xi' + \sigma', \cdot))(\kappa(\xi) - \kappa(\sigma'))\} \tag{34}$$

where C denotes the cosine transform:

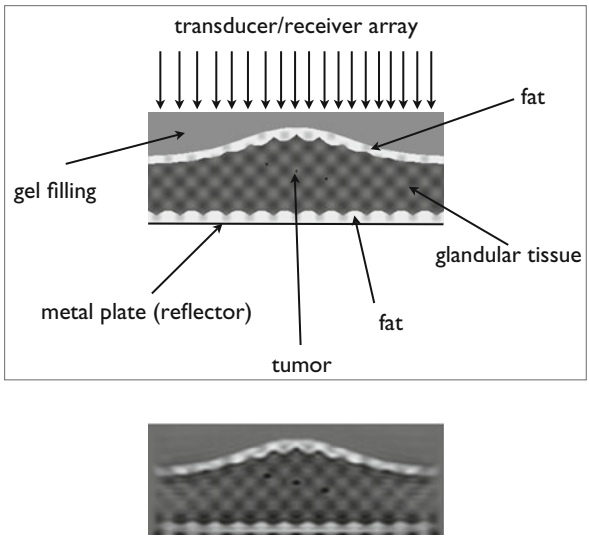


Fig. 10 *Top*: Principle of CARI. *Bottom*: Reconstruction from CARI-data with 20 Kaczmarz sweeps

$$(Cf)(\xi) = \int_0^D \cos(\xi x) f(x) dx.$$

Equation (34) is the solution in the Born approximation of the inverse scattering problem with a reflector. The essential difference to the solution without a reflector, as given in (9), is that the argument $\kappa(\rho') - \kappa(\sigma')$ shows up, indicating that small frequencies are involved.

In [43] it is shown that (34) in fact permits the determination of the Fourier transform of f for frequencies down to 0. In [45] it is shown what can be achieved by a combination of reflectors with the Kaczmarz method in mammography [5, 58] see Fig. 10. Again we use 19 sources at the top boundary. The relaxation parameter α is 3×10^7. We work with a source pulse whose signature is uniform from 50 to 500 kHz. Without the mirror the results would be completely unsatisfactory.

8 Direct Methods

In this section we mention some non-iterative methods which provide exact solutions to the nonlinear problem.

Boundary Control

The boundary control method for inverse problems of hyperbolic differential equation has been developed in [1]. We follow [55] and start out from the formulation

(1)–(3) of the inverse problem. Let u_h be the solution of (1) with $\partial u/\partial \nu = h$ on $\Gamma \times (0,T)$ and $a^+(t) = (a(t) + a(2T-t))/2$ for any function a. Let for functions h, g on $\Gamma \times (0,T)$

$$[h,g] = \int_\Omega c^{-2}(x) u_{h_t}(x,T) u_{g_t}(x,T) dx,$$

$$[h,g]_1 = \int_\Omega \nabla u_h(x,T) \cdot \nabla u_g(x,T) dx.$$

The key to the boundary control method are the relations

$$[h,g] = \int_\Gamma \int_0^T (u_g^+ h - g^+ u_h) dx dt, \tag{35}$$

$$[h,g]_1 = \int_\Gamma \int_0^T (u_{g_t}^+ h + g^+ u_{h_t}) dx dt. \tag{36}$$

Note that the right hand sides of (35), (36) can be evaluated without knowing c. All we need to know is u_h, u_g on $\Gamma \times (0,2T)$ for the functions h, g, and this is just the data for the inverse problem.

Now let φ be any harmonic function in Ω. We determine a function h^φ on $\Gamma \times (0,T)$ such that $u_{h^\varphi} = \varphi$. Let $h_s(x,t) = q(t)\delta(x-s)$ for each source s on Γ. We try the Ansatz

$$h^\varphi = \sum_s a_s h_s$$

with certain coefficients a_s, i.e.,

$$\varphi(x) = \sum_s a_s u_{h_s}(x,T).$$

The coefficients a_s may be determined from the equations

$$\int_\Omega \nabla\varphi(x) \cdot \nabla u_{h_w}(x,T) dx = \sum_s a_s \int_\Omega \nabla u_{h_s}(x,T) \cdot \nabla u_{h_w}(x,T) dx$$

which, by the Gauss integral theorem can be written as

$$\int_\Gamma \frac{\partial \varphi}{\partial \nu}(x) u_{h_w}(x,T) dx = \sum_s a_s [h_w, h_s]_1 \tag{37}$$

The left hand side and the coefficients on the right hand side are determined by the data. Solving (37) for a_s we get the function h^φ. We do this for many harmonic functions φ, ψ. By (35) we then can compute the numbers

$$\int_{\Omega} c^{-2}\varphi\psi dx = [h^{\varphi}, h^{\psi}].$$

Since the products of harmonic functions are complete in $L_2(\Omega)$, this determines c.

Inverse Scattering

The inverse problem of ultrasound tomography is a typical case of inverse scattering. Uniqueness questions have been dealt with in [37, 52]. Since the methods in these papers are, in principle, constructive, numerical methods can be based on these methods, see [7].

9 Conclusion

The mathematics of sonic imaging is well understood, at least for the linearized problem with constant background. Reasonably efficient iterative methods for the fully nonlinear problems are available for transmission imaging. Time domain as well as frequency domain methods are being used. In reflection imaging the main problem is the lack of low frequencies in the source pulse. It is not clear at present if this problem ever finds a satisfactory solution.

Cross-References

▸ Inverse Scattering
▸ Iterative Solution Methods

References

1. Belishev, M.I., Gotlib, V.Yu: Dynamical variant of the BC-method: theory and numerical testing. J. Inv. Ill-Posed Prob. **7**, 221–240 (1999)
2. Bingham, K., Kurylev, Y., Lassas, M., Siltanen, S.: Iterative time-reversal control for inverse problems. Inverse Prob. Imaging **2**, 63–81 (2008)
3. Borcea, L., Papanicolaou, G., Tsogka, C.: Theory and application of time reversal and interferometric imaging. Inverse Prob. **19**, S139–S164 (2003)
4. Borup, D.T., Johnson, S.A., Kim, W.W., Berggren, M.J.: Nonperturbative diffraction tomography via Gauss–Newton iteration applied to the scattering integral equation. Ultrason. Imaging **14**, 69–85 (1992)
5. Bounheim, A., et al.: FETD simulation of wave propagation modeling the Cari breast sonography. In: Kumar et al. (eds.) Lecture Notes in Computer Science, vol. 2668, pp. 705–714. Springer, New York (2003)
6. Burger, M., Kaltenbacher, B.: Regularizing Newton–Kaczmarz methods for nonlinear ill-posed problems. SIAM J. Numer. Anal. **44**,153–182 (2006)

7. Burov, A.V., Morozov, S.A., Rumyantseva, O.D.: Reconstruction of fine-scale structure of acoustical scatterer on large-scale contrast background. In: Acoustical Imaging, vol. 26, pp. 231–238. Kluwer Academic/Plenum, New York (2006)
8. Chen, Yu: Inverse scattering via Heisenberg's uncertainty principle. Inverse Prob. **13**, 253–282 (1997)
9. Claerbout, J.F.: Fundamentals of Geophysical Data Processing. McGraw-Hill, New York (1976)
10. Claerbout, J.F.: Imaging the Earth's Interior. Blackwell, Oxford (1985)
11. Clemmow, P.: The Plane Wave Spectrum Representation of Electromagnetic Fields. Oxford University Press, New York (1996)
12. Colton, D., Kress, R.: Inverse Acoustic and Electromagnetic Scattering Theory, 3rd edn. Springer, New York (2013)
13. DeHoop, M.: Microlocal analysis of seismic inverse scattering. In: Uhlmann, G. (ed.) Inside Out, pp. 219–296. Cambridge University Press, Cambridge (2003)
14. Devaney, A.J.: A filtered backpropagation algorithm for diffraction tomography. Ultrason. Imaging **4**, 336–350 (1982)
15. Devaney, A.J.: Mathematical Foundations of Imaging, Tomography and Wavefield Inversion. Cambridge University Press, Cambridge (2012)
16. Dierkes, T., Dorn, O., Natterer, F., Palamodov, V., Sielschott, H.: Fréchet derivatives for some bilinear inverse problems. SIAM J. Appl. Math. **62**, 2092–2113 (2002)
17. Divakar, S.: 3D Ultrasound Medical Imaging from Reflection Data. Master's thesis, Audio-visual Communications Laboratory (LCAV), École Polytechnique Fédérale de Lausanne (2013)
18. Dorn, O., Lesselier, D.: Level set methods for inverse scattering. Inverse Prob. **22**, R67–R131 (2006)
19. Dragoset, B., Gabitsch, J.: Introduction to this special section: low-frequency seismic. Lead. Edge **26**, 34–36 (2007)
20. Duric, N., et al.: Development of ultrasound tomography for breast imaging: technical assessment. Med. Phys. **32**, 1375–1386 (2005)
21. Dussik, K.T.: Über die Möglichkeit hochfrequente mechanische Schwingungen als diagnostische Hilfsmittel zu verwenden. Z. f. d. ges. Neurol. u. Psychiat. **174**, 143 (1942)
22. Gauthier, O., Virieux, J., Tarantola, A.: Two-dimensional nonlinear inversion of seismic waveforms: numerical results. Geophysics **51**, 1387–1403 (1986)
23. Goncharsky, A.V., Romanov, S.Y: Supercomputer technologies in inverse problems of ultrasound tomography. Inverse Prob. **29**, 075004 (2013)
24. Gutman, S., Klibanov, M.: Iterative method for multi-dimensional inverse scattering problems at fixed frequencies. Inverse Prob. **10**, 573–599 (1994)
25. Herman, G.T.: Image Reconstruction from Projections. Academic, London (1980)
26. Hesse, M.C., Salehi, L., Schmitz, G.: Nonlinear simultaneous reconstruction of inhomogeneous compressibility and mass density distributions in unidirectional pulse-echo ultrasound imaging. Phys. Med. Biol. **58**, 6163–6178 (2013)
27. Hohage, T.: On the numerical solution of a three-dimensional inverse medium scattering problem. Inverse Prob. **17**(2001), 1743–1763 (2001)
28. Hristova Y., Kuchment, P., Nguyen, L.: Reconstruction and time reversal in thermoacoustic tomography in acoustically homogeneous and inhomogeneous media. Inverse Prob. **24**, 055006 (2008)
29. Isakov, V.: Inverse Problems for Partial Differential Equations. Applied Mathematical Sciences, vol. 127. Springer, New York (1998)
30. Jannane, M., et al.: Wavelengths of earth structures that can be resolved from seismic reflection data. Geophysics **54**, 906–910 (1989)
31. Jonas, P., Louis, A.K.: Phase contrast tomography using holographic measurements. Inverse Prob. **20**, 75–102 (2004)
32. Kac, A.C., Slaney, M.: Principle of Computerized Tomographic Imaging. IEEE Press, New York (1988)

33. Kleinman, R.E., van den Berg, P.M.: A modified gradient method for two-dimensional problems in tomography. J. Comput. Appl. Math. **42**, 17–36 (1992)
34. Kosloff, D., Baysal, E.: Migration with the full acoustic wave equation. Geophysics **48**, 677–687 (1983)
35. Mora, P.: Inversion = migration + tomography. Geophysics **54**, 1575–1586 (1989)
36. Morgan, S.P.: General solution of the Luneburg lens problem. J. Appl. Phys. **29**, 1358–1368 (1958)
37. Nachman, A.I.: Reconstructions from boundary measurements. Ann. Math. **128**, 531–576 (1988)
38. Natterer, F.: Numerical solution of bilinear inverse problems, Technical Report 19/96-N. Fachbereich Mathematik und Informatik der Universität Münster (1996)
39. Natterer, F.: An algorithm for 3D ultrasound tomography. In: Chavent, G., Sabatier, P. (eds.) Inverse Problems of Wave Propagation. Lecture Notes in Physics, pp. 216–225. Springer, New York (1997)
40. Natterer, F.: An algorithm for the fully nonlinear inverse scattering problem at fixed frequency. J. Comput. Acoust. **9**, 935–940 (2001)
41. Natterer, F.: An error bound for the Born approximation. Inverse Prob. **20**, 447–452 (2004)
42. Natterer, F.: Ultrasound tomography with fixed linear arrays of transducers. In: Proceedings of the Interdisciplinary Workshop on Mathematical Methods in Biomedical Imaging and Intensity-Modulated Radiation Therapy (IMRT), Pisa (2007)
43. Natterer, F.: Reflectors in wave equation imaging. Wave Motion **45**, 776–784 (2008)
44. Natterer, F.: Acoustic imaging in 3D. In: Censor, Y., Jiang, M., Wang, G. (eds.) Biomedical Mathematics: Promising Directions in Imaging, Therapy Planning, and Inverse Problems. Medical Physics, Madison (2009)
45. Natterer, F.: Ultrasound mammography with a mirror. Phys. Med. Biol. **55**, N275–N299 (2010)
46. Natterer, F.: Reflection imaging without low frequencies. Inverse Prob. **27**, 035011 (2011)
47. Natterer, F.: Reflection imaging of layered media without using low frequencies. Inverse Prob. **29**, 035001 (2013)
48. Natterer, F., Klyubina, O.: Initial value techniques for the Helmholtz and Maxwell equations. J. Comput. Math. **25**, 368–373 (2007)
49. Natterer, F., Wübbeling, F.: A propagation-backpropagation method for ultrasound tomography. Inverse Prob. **11**, 1225–1232 (1995)
50. Natterer, F., Wübbeling, F.: Mathematical Methods in Image Reconstruction. SIAM, Philadelphia (2001)
51. Natterer, F., Wübbeling, F.: Marching schemes for inverse acoustic scattering problems. Numer. Math. **100**, 697–710 (2005)
52. Novikov, R.G.: The $\bar{\partial}$-approach to approximate inverse scattering at fixed energy in three dimensions. Int. Math. Res. Pap. **6**, 287–349 (2005)
53. Palamodov, V.P.: Stability of diffraction tomography and a nonlinear "basic theorem". J. Anal. Math **91**, 247–286 (2003)
54. Palamodov, V.: Inverse scattering as nonlinear tomography. Wave Motion **47**, 635–640 (2010)
55. Pestov, L., Bolgova V., Kazarina, O.: Numerical recovering of a density by the BC-method. Inverse Prob. Imaging **2**, 703–712 (2008)
56. Pintavirooj, C., Sangworasil, M.: Ultrasonic diffraction tomography. Int. J. Appl. Biomed. Eng. **1**, 34–40 (2008)
57. Prada, C., Kerbrat, E., Cassereau, D., Fink, M.: Time reversal techniques in ultrasonic nondestructive testing of scattering media. Special section on electromagnetic and ultrasonic nondestructive evaluation. Inverse Prob. **18**, 1761–1773 (2002)
58. Richter, K.: Clinical amplitude/velocity reconstructive imaging (CARI) - a new sonographic method for detecting breast lesions. Br. J. Radiol. **68**, 375–384 (1995)
59. Sandberg, K., Beylkin, G.: Full wave-equation depth extrapolation for migration. Geophysics **74**, WCA121–WCA128 (2009)
60. Santosa, F., Symes, W.W.: An Analysis of Least-Squares Velocity Inversion. Geophysical Monographs Series, vol. 4. Society of Exploration Geophysics, Tulsa (1989)

61. Sielschott, H.: Rückpropagationsverfahren für die Wellengleichung in bewegtem Medium. Ph.D. thesis, Preprints Angewandte Mathematik und Informatik, 15/00-N, Münster, Germany (2000). www.math.uni-muenster.de/num/Preprints/2000/sielsch
62. Sirgue, L., Pratt, R.G.: Efficient waveform inversion and imaging: a strategy for selecting temporal frequencies. Geophysics **69**, 231–248 (2004)
63. Symes, W.W.: The seismic reflection inverse problem. Inverse Prob. **25**, 123008, 39 pp. (2009)
64. Wu, R., Toksöz, M.N.: Diffraction tomography and multisource holography applied to seismic imaging. Geophysics **52**, 11–25 (1987)

Imaging in Random Media

Liliana Borcea

Contents

1 Introduction 1280
2 Main Text 1282
3 Basic Imaging 1282
The Forward Model 1283
Least Squares Inversion 1288
The Normal Operator and the Time Reversal Process 1292
Imaging in Smooth and Known Media 1295
Robustness to Additive Noise 1299
4 Challenges of Imaging in Complex Media 1302
Cluttered Media 1303
The Random Model 1303
Time Reversal Is Not Imaging 1305
5 The Random Travel Time Model of Wave Propagation 1306
Long Range Scaling and Gaussian Statistics 1308
Statistical Moments 1309
6 Setup for Imaging 1313
7 Migration Imaging 1315
The Expectation 1316
The SNR 1318
8 CINT Imaging 1319
Analysis of the Cross-Correlations for a Point Source 1321
Resolution Analysis of the CINT Imaging Function 1324
CINT Images for Passive Arrays as the Smoothed Wigner Transform 1327
CINT Imaging with Active Arrays 1331
9 Appendix 1: Second Moments of the Random Travel Time 1335
10 Appendix 2: Second Moments of the Local Cross-Correlations 1337
11 Conclusion 1338
Cross-References 1339
References 1339

L. Borcea (✉)
Department of Mathematics, University of Michigan, Ann Arbor, MI, USA
e-mail: borcea@umich.edu

© Springer Science+Business Media New York 2015

O. Scherzer (ed.), *Handbook of Mathematical Methods in Imaging*,
DOI 10.1007/978-1-4939-0790-8_41

Abstract

We give a self-contained presentation of coherent array imaging in random media, which are mathematical models of media with uncertain small-scale features (inhomogeneities). We describe the challenges of imaging in random media and discuss the coherent interferometric (CINT) imaging approach. It is designed to image with partially coherent waves, so it works at distances that do not exceed a transport mean-free path. The waves are incoherent when they travel longer distances, due to strong cumulative scattering by the inhomogeneities, and coherent imaging becomes impossible. In this article we base the presentation of coherent imaging on a simple geometrical optics model of wave propagation with randomly perturbed travel time. The model captures the canonical form of the second statistical moments of the wave field, which describe the loss of coherence and decorrelation of the waves due to scattering in random media. We use it to give an explicit resolution analysis of CINT which includes the assessment of statistical stability of the images.

1 Introduction

Sensor array imaging is an important technology in a variety of applications such as medical ultrasound, nondestructive evaluation of materials, underwater acoustics, geophysical prospecting, radar, and elsewhere. In general terms, it seeks to estimate wave sources or reflecting structures in a medium using measurements at the sensors, which are devices that transform one form of energy into another. In radar the sensors are antennas that convert electromagnetic waves to/from electrical signals. In underwater acoustics they are hydrophones that convert changes in water pressure to electrical signals, and so on.

The array consists of many sensors which are placed close together on a measurement surface, so that they behave like a collective entity. It occupies a bounded set, the *array aperture*, which we denote by $\mathcal{A}$. In problems like synthetic aperture radar [16], there isn't an actual array. Instead, one or more sensors mounted on a moving platform span a long path, the synthetic aperture. To fix ideas we assume throughout an actual array, but most of the results presented here extend to problems with synthetic apertures.

When the sensors are just receivers, we say that the array is *passive*. The sources of waves are typically far away from the passive array, and the problem is to determine them from the measurements at the receivers. *Active* arrays consist of sensors that are sources and receivers. They probe the medium with waves emitted by one or many sources and receive the echoes from the unknown reflecting structures.

The recordings at the receivers are called the *array data*. We are concerned with *coherent imaging* systems, meaning that the data are time-resolved measurements of the waves. Equivalently, in the Fourier domain, the data are measurements of the amplitude and the phase of the Fourier components of the waves, for all

the frequencies in the bandwidth. Incoherent imaging systems like in infrared or optical tomography measure the net intensity of the waves, which are typically single frequency or narrow-band. They image at distances that are larger than a transport mean-free path, a length scale which marks the onset of diffusion in random media [29, 35]. They involve very different data processing and give worse image resolution than the coherent ones.

Imaging is a simplified inverse problem for the wave equation. One example of an inverse problem is to determine the coefficients in the wave equation, like the wave speed, from measurements of the solution at the sensor locations. This is a nonlinear and ill-posed problem that is difficult to solve [30, 33, 34]. Imaging seeks less information about the medium, such as the loci of jumps of the wave velocity, which model the location of reflecting structures. It is an easier problem as long as we know the source of waves and how the waves propagate.

An imaging system is useful only if it is robust to uncertainties in the problem. We may think of uncertainties as "noise," but not all noise is the same. Some is easier to deal with, like additive measurement noise considered in most of the applied imaging literature. However, uncertainties in the model of wave propagation are much harder to mitigate. We study the effect of such uncertainties arising in imaging in complex media with numerous small inhomogeneities, i.e., with microstructure. This is unknown in practice and cannot be estimated from the necessarily band-limited data as part of the imaging process. The microstructure may not even be interesting in applications. However, it cannot be ignored because although one inhomogeneity by itself is negligible, there are many of them and their cumulative wave scattering effect builds up as the waves travel in the medium.

We deal with the unknown microstructure by modeling it with random processes, and thus study *imaging in random media*. We describe first the forward problem, that is the mapping from the random microstructure to the solution of the wave equation at the location of the sensors. This solution is a random field. The data are one realization of this field, corresponding to one realization of the random medium. The challenge is to find imaging methods that are robust, i.e., that produce images which do not change significantly with the realization of the random medium. The robustness is called *statistical stability*.

To analyze imaging in random media, we study the statistics of the wave field at the sensors. Specifically, we calculate the mean field, which we call the *coherent field*, and its random fluctuations, the *incoherent field*. When the fluctuations are small, the *data is coherent* and imaging is easier. When the fluctuations dominate the mean field, the *data is incoherent*. This is a typical regime in applications like diffusion tomography. We look at a regime between these two, where the *data is partially coherent*. The cumulative scattering effects are important in this regime, and mathematics allows us to understand them under simplifying assumptions like: the inhomogeneities are weak scatterers; there is separation of scales; the microstructure decorrelates rapidly with distance. Recent results [22, 25] have relaxed the latter assumption to media with long range correlation, and the case of strong inhomogeneities can be handled in some media [21].

The analysis of wave propagation in random media is quite involved, and it is specific to the scaling regime. There is no universal treatment of the problem, no one fits all regime. The detailed quantification of cumulative scattering effects and the calculation of important scales such as the scattering mean-free path and the transport mean-free path depend on the assumptions on the microstructure and the frequency range [35]. However, there are canonical effects such as the loss of coherence and decorrelation of the waves that can be given a generic description, via the second statistical moments of the random wave field. We describe them with the simplest model of wave propagation in random media, which accounts only for random wavefront distortions. It is a geometrical optics approximation of the solution of the wave equation, with randomly perturbed travel time. We use this model for the purpose of a self-contained analysis of coherent imaging in random media. The results extend to more general wave scattering regimes, as described in [3, 7–9, 18, 27].

2 Main Text

The presentation is organized as follows: We begin in Sect. 3 with the basics of imaging. We formulate the data model and the reverse time migration imaging function as an approximate least squares solution of the inverse problem. Migration is the basic imaging method in many applications. We describe in Sect. 4 the challenges of imaging in complex media. The analysis of imaging in such media is in Sects. 7 and 8. It uses the random travel time model described in Sect. 5 and the setup given in Sect. 6. We show with detailed calculations in Sect. 7 how migration imaging fails in complex media with significant cumulative wave scattering by the microstructure. Then we analyze in Sect. 8 the coherent interferometric (CINT) approach, which is designed to image with partially coherent waves. We give a detailed resolution analysis of this method, including the assessment of its statistical stability and illustrate the results with numerical simulations.

3 Basic Imaging

In this section we present a mathematical formulation of imaging with sensor arrays. We begin in section "The Forward Model" with the description of the model of the data recorded by passive and active arrays. We call it the *forward model* because it maps the source distribution or the reflectivity, the unknowns in the imaging problem, to the wave field measured by the sensors. In general there is no explicit inverse mapping from the array data to the unknown source density or reflectivity. We show in section "Least Squares Inversion" how we can state the imaging problem in variational, *least squares data fit* form. The mathematical model of the *time reversal* process can be viewed as an approximation of the solution of the least squares problem, as described in section "The Normal Operator and the Time Reversal Process". It can be computed and used for imaging only when we

know the medium through which the waves propagate. We discuss its refocusing in smooth media and robustness to additive noise in section "Imaging in Smooth and Known Media."

The Forward Model

We base the forward model of the array data on the scalar wave equation for the acoustic pressure $p(t,\vec{\mathbf{x}})$

$$\frac{1}{c^2(\vec{\mathbf{x}})}\frac{\partial^2 p(t,\vec{\mathbf{x}})}{\partial t^2} - \Delta p(t,\vec{\mathbf{x}}) = F(t,\vec{\mathbf{x}}), \quad \vec{\mathbf{x}} \in \mathbb{R}^3, \ \ t > 0, \tag{1}$$

where $c(\vec{\mathbf{x}})$ is the wave speed and $F(t,\vec{\mathbf{x}})$ the source density. We call $p(t,\vec{\mathbf{x}})$ the wave field and relate it to the source using Duhamel's principle

$$p(t,\vec{\mathbf{x}}) = \int_0^t ds \int_{\mathbb{R}^n} d\vec{\mathbf{y}}\, F(s,\vec{\mathbf{y}})G(t-s,\vec{\mathbf{y}},\vec{\mathbf{x}}) = \int_{\mathbb{R}^n} d\vec{\mathbf{y}}\, F(t,\vec{\mathbf{y}}) \star_t G(t,\vec{\mathbf{y}},\vec{\mathbf{x}}). \tag{2}$$

Here $G(t,\vec{\mathbf{x}},\vec{\mathbf{y}})$ is the causal Green's function, the solution in the sense of distributions of

$$\frac{1}{c^2(\vec{\mathbf{x}})}\frac{\partial^2 G(t,\vec{\mathbf{x}},\vec{\mathbf{y}})}{\partial t^2} - \Delta_{\vec{\mathbf{x}}} G(t,\vec{\mathbf{x}},\vec{\mathbf{y}}) = \delta(\vec{\mathbf{x}}-\vec{\mathbf{y}})\delta(t), \quad \vec{\mathbf{x}},\vec{\mathbf{y}} \in \mathbb{R}^n, \ \ t > 0, \tag{3}$$

$$G(t,\vec{\mathbf{x}},\vec{\mathbf{y}}) = 0, \quad t < 0, \tag{4}$$

where $\Delta_{\vec{\mathbf{x}}}$ is the Laplace operator in the $\vec{\mathbf{x}}$ variable, and δ is the Dirac distribution. The second equality in (2) is because both the source and Green's function are causal (supported at time $t > 0$), and we can write

$$\int_0^t ds\, F(s,\vec{\mathbf{y}})G(t-s,\vec{\mathbf{x}},\vec{\mathbf{y}}) = \int_{-\infty}^{\infty} ds\, F(s,\vec{\mathbf{y}})G(t-s,\vec{\mathbf{y}},\vec{\mathbf{x}}) = F(t,\vec{\mathbf{y}}) \star_t G(t,\vec{\mathbf{x}},\vec{\mathbf{y}}).$$

Convolutions are not convenient for the analysis, so we often work in the Fourier (frequency) domain. The Fourier transform of the wave field is defined as

$$\hat{p}(\omega,\vec{\mathbf{x}}) = \int_{-\infty}^{\infty} dt\, p(t,\vec{\mathbf{x}})e^{i\omega t}, \tag{5}$$

and the inverse transform is

$$p(t,\vec{\mathbf{x}}) = \int_{-\infty}^{\infty} \frac{d\omega}{2\pi}\, \hat{p}(\omega,\vec{\mathbf{x}})e^{-i\omega t}. \tag{6}$$

The dual variable to time t in these transformations, denoted by ω, is the angular frequency (measured in radians per second). It is related to the frequency (measured in Hz) by a factor of 2π. We also relate ω to the *wavenumber* k and the *wavelength* λ using a reference wave speed scale c_o,

$$k = \frac{\omega}{c_o} = \frac{2\pi}{\lambda}. \tag{7}$$

The advantage of working in the Fourier domain is that the convolution in (2) becomes a product

$$\hat{p}(\omega, \vec{\mathbf{x}}) = \int_{\mathbb{R}^n} d\vec{\mathbf{y}}\, \hat{F}(\omega, \vec{\mathbf{y}}) \hat{G}(\omega, \vec{\mathbf{y}}, \vec{\mathbf{x}}), \tag{8}$$

with $\hat{G}(\omega, \vec{\mathbf{y}}, \vec{\mathbf{x}})$ the Fourier transform of the causal Green's function. This is the same as the outgoing Green's function of the Helmholtz equation.

Passive Array Data Model

When imaging with passive arrays, $F(t, \vec{\mathbf{x}})$ is unknown. We assume for simplicity that it has the separable form

$$F(t, \vec{\mathbf{x}}) = f(t)\rho(\vec{\mathbf{x}}), \tag{9}$$

where the same signal $f(t)$ is emitted from all the points in the support of the source density $\rho(\vec{\mathbf{x}})$. In some applications like synthetic aperture radar, $f(t)$ is a complex valued and long (chirped) signal. But in other applications $f(t)$ is a function of small temporal support, a pulse. We assume henceforth such a pulse and model $f(t)$ as a base-band pulse $f_B(t)$ modulated by an harmonic signal at carrier (central) frequency $\omega_o/(2\pi)$,

$$f(t) = e^{i\omega_o t} f_B(t). \tag{10}$$

The terminology becomes clear in the frequency domain, where

$$\hat{f}(\omega) = \int_{-\infty}^{\infty} d\omega\, f(t) e^{i\omega t} = \hat{f}_B(\omega - \omega_o). \tag{11}$$

The signal $f_B(t)$ is called base-band pulse because its Fourier transform $\hat{f}_B(\omega)$ is supported at $\omega \in [-\pi B, \pi B]$. Then, the frequencies $\omega/2\pi$ in the support of $\hat{f}(\omega)$ lie in the interval centered at $\omega_o/(2\pi)$ of bandwidth B.

The N_r receivers located at $\vec{\mathbf{x}}_r$ in the aperture $\mathcal{A}$ of the array measure the data

$$D(t) = \left\{ d(t, \vec{\mathbf{x}}_r) \right\}, \quad t \in (0, T], \quad \vec{\mathbf{x}}_r \in \mathcal{A}, \quad r = 1, \ldots, N_r, \tag{12}$$

over the time duration T. Equivalently, in the Fourier domain we have

$$\hat{D}(\omega) = \left\{ \hat{d}(\omega, \vec{\mathbf{x}}_r) \right\}, \quad |\omega - \omega_o| \leq \pi B, \ \ \vec{\mathbf{x}}_r \in \mathcal{A}, \ r = 1, \ldots, N_r. \tag{13}$$

The goal of imaging is to estimate from these data the support of ρ, which is contained in a compact set $I_W \subset \mathbb{R}^n$ called the *imaging window*.

We use a terminology borrowed from the seismic imaging literature, and refer to $d(t, \vec{\mathbf{x}}_r)$ as *time traces*, to emphasize that they are time recordings. Their model comes from (2),

$$d(t, \vec{\mathbf{x}}_r) = 1_{(0,T]}(t) \left[\int_{-\infty}^{\infty} \frac{d\omega}{2\pi} \int_{I_W} d\vec{\mathbf{y}} \, \hat{F}(\omega, \vec{\mathbf{y}}) \hat{G}(\omega, \vec{\mathbf{x}}_r, \vec{\mathbf{y}}) \, e^{-i\omega t} + \mathcal{N}_r(t) \right], \tag{14}$$

where we added the noise $\mathcal{N}_r(t)$ indexed by the receiver. We assume as is typical that $\mathcal{N}_r(t)$ is white noise, i.e., a real-valued stationary Gaussian generalized process with zero expectation, covariance equal to the delta distribution, and independent over the receivers

$$\mathbb{E}\left[\mathcal{N}_r(t)\right] = 0, \qquad \mathbb{E}\left[\mathcal{N}_r(t)\mathcal{N}_{r'}(t')\right] = \sigma_{\mathcal{N}}^2 \delta(t - t') \delta_{r,r'}. \tag{15}$$

Here $\mathbb{E}$ denotes statistical expectation; $\delta_{r,r'}$ is the Kronecker delta symbol; the parameter $\sigma_{\mathcal{N}}$ scales the noise level and the time window $1_{(0,T]}(t)$ is equal to one when $t \in (0, T]$ and zero otherwise. The waves are transient and thus can be captured for large enough T. We assume such a T so we can ignore the window $1_{(0,T]}(t)$ in the first term of (14), and obtain the following relation in the Fourier domain,

$$\hat{d}(\omega, \vec{\mathbf{x}}_r) \approx \int_{I_W} d\vec{\mathbf{y}} \, \hat{F}(\omega, \vec{\mathbf{y}}) \hat{G}(\omega, \vec{\mathbf{x}}_r, \vec{\mathbf{y}}) + \hat{\mathcal{N}}_r^T(\omega), \quad \hat{\mathcal{N}}_r^T(\omega) = \int_0^T dt \, \mathcal{N}_r(t) e^{i\omega t}. \tag{16}$$

The forward map $\mathcal{M}$ takes the unknown source density ρ to the data space $\mathbb{D}$. We show in the next section how to formulate the inverse problem as an optimization that computes an "approximate inverse" of $\mathcal{M}$ in some sense. This is not an actual inverse which may not even exist. For optimization, it is convenient to work in function spaces with inner products. Thus we assume that $\rho(\vec{\mathbf{y}})$ lies in $L^2(I_W)$, the Hilbert space of real valued and square integrable functions supported in I_W, with inner product

$$(\rho, \eta) = \int_{I_W} d\vec{\mathbf{y}} \rho(\vec{\mathbf{y}}) \eta(\vec{\mathbf{y}}), \quad \forall \ \rho, \eta \in L^2(I_W), \tag{17}$$

and norm $\|\rho\|_2 = \sqrt{(\rho, \rho)}$. The data lies in the vector space of functions of finite energy

$$\mathbb{D} = \left\{ \hat{d}(\omega, \vec{\mathbf{x}}_r) \in \mathbb{C}, \text{ s.t. } \sum_{r=1}^{N_r} \int_{|\omega-\omega_o|\le \pi B} d\omega \left| \hat{d}(\omega, \vec{\mathbf{x}}_r) \right|^2 < \infty \right\}, \tag{18}$$

endowed with the complex inner product

$$\left\langle \hat{d}, \hat{g} \right\rangle = \int_{|\omega-\omega_o|\le \pi B} d\omega \sum_{r=1}^{N_r} \overline{\hat{d}(\omega, \vec{\mathbf{x}}_r)} \hat{g}(\omega, \vec{\mathbf{x}}_r), \quad \forall\ \hat{d}, \hat{g} \in \mathbb{D}, \tag{19}$$

and the norm $\|\hat{d}\|_{\mathbb{D}} = \sqrt{\left\langle \hat{d}, \hat{d} \right\rangle}$.

The forward map $\mathcal{M} : L^2(I_W) \to \mathbb{D}$ is the linear operator defined by

$$[\mathcal{M}\rho](\omega, \vec{\mathbf{x}}_r) = \hat{f}(\omega) \int_{I_W} d\vec{\mathbf{y}}\, \rho(\vec{\mathbf{y}}) \hat{G}(\omega, \vec{\mathbf{x}}_r, \vec{\mathbf{y}}). \tag{20}$$

It assumes that we know the pulse and the wave speed $c(\vec{\mathbf{x}})$, i.e., the Green's function. The latter assumption turns out to be critical, and it is not satisfied in complex media with microstructure.

Active Array Data Model

To state the model for active arrays, we redefine the coefficient in the wave equation as

$$\frac{1}{c^2(\vec{\mathbf{x}})} \rightsquigarrow \frac{1}{c^2(\vec{\mathbf{x}})} + \frac{\rho(\vec{\mathbf{x}})}{c_o^2}, \tag{21}$$

where $c(\vec{\mathbf{x}})$ is the *assumed known* wave speed in the medium which hosts the perturbation (reflectivity) $\rho(\vec{\mathbf{x}})$ that we wish to estimate. It is supported in a compact set $I_W \subset \mathbb{R}^n$ that defines the imaging region. The name *reflectivity* suggests that we expect it to cause reflected waves (echoes) that can be measured at the array. This is relevant for most imaging setups where the array lies on one side of the medium. For example, in exploration geophysics, the receivers lie on the surface of the earth and do not see waves that interact with the subsurface unless ρ reflects them.

Suppose that the excitation comes from a source in the array, idealized as a point at $\vec{\mathbf{x}}_s \in \mathcal{A}$,

$$F(t, \vec{\mathbf{x}}) = f(t)\delta(\vec{\mathbf{x}} - \vec{\mathbf{x}}_s). \tag{22}$$

The array has $N_s \ge 1$ sources which may emit simultaneously, but we assume here that the excitation is with one source at a time, and we emphasize with the notation $p(t, \vec{\mathbf{x}}, \vec{\mathbf{x}}_s)$ the dependence of the wave field on the source location. It satisfies

$$\frac{1}{c^2(\vec{\mathbf{x}})}\frac{\partial^2 p(t,\vec{\mathbf{x}},\vec{\mathbf{x}}_s)}{\partial t^2} - \Delta_{\vec{\mathbf{x}}}\, p(t,\vec{\mathbf{x}},\vec{\mathbf{x}}_s) = -\frac{\rho(\vec{\mathbf{x}})}{c_o^2}\frac{\partial^2 p(t,\vec{\mathbf{x}},\vec{\mathbf{x}}_s)}{\partial t^2} + f(t)\delta(\vec{\mathbf{x}}-\vec{\mathbf{x}}_s), t>0, \tag{23}$$

with initial condition

$$p(0,\vec{\mathbf{x}},\vec{\mathbf{x}}_s) = \frac{\partial}{\partial t}p(0,\vec{\mathbf{x}},\vec{\mathbf{x}}_s) = 0, \tag{24}$$

and we can write it in terms of the causal Green's function using Duhamel's principle

$$p(t,\vec{\mathbf{x}},\vec{\mathbf{x}}_s) = f(t) \star_t G(t,\vec{\mathbf{x}},\vec{\mathbf{x}}_s) - \int_{I_W} d\vec{\mathbf{y}}\, \frac{\rho(\vec{\mathbf{y}})}{c_o^2}\frac{\partial^2 p(t,\vec{\mathbf{y}},\vec{\mathbf{x}}_s)}{\partial t^2} \star_t G(t,\vec{\mathbf{x}},\vec{\mathbf{y}}). \tag{25}$$

In the forward model we evaluate (25) at the receiver locations $\vec{\mathbf{x}} = \vec{\mathbf{x}}_r \in \mathcal{A}$, for $r = 1,\ldots,N_r$. (The sources and receivers may be collocated). The first term $f(t) \star_t G(t,\vec{\mathbf{x}}_r,\vec{\mathbf{x}}_s)$ is the direct wave that does not interact with the reflectivity. We can remove it by time windowing, assuming that it arrives much earlier than the echoes from the medium. Thus, we redefine the origin of time of the measurements after the direct arrival of the wave from $\vec{\mathbf{x}}_s$, and let henceforth $p(t,\vec{\mathbf{x}}_r,\vec{\mathbf{x}}_s)$ model the echoes from the medium for $t \in (0,T]$. The second term in (25) depends nonlinearly on ρ. We use its linearization, known as the *Born approximation*, as is typical in imaging. The approximation holds when ρ has small amplitude and/or small support, although numerical experiments suggest that it may have a wider range of validity. We refer to [32] for its analysis in media with general bounded background wave speed $c(\vec{\mathbf{x}})$.

The Born data model is

$$d(t,\vec{\mathbf{x}}_r,\vec{\mathbf{x}}_s) \approx \int_{-\infty}^{\infty}\frac{d\omega}{2\pi}k^2\hat{f}(\omega)\int_{I_W} d\vec{\mathbf{y}}\rho(\vec{\mathbf{y}})\hat{G}(\omega,\vec{\mathbf{x}}_r,\vec{\mathbf{y}})\hat{G}(\omega,\vec{\mathbf{y}},\vec{\mathbf{x}}_s)e^{-i\omega t} + 1_{(0,T]}(t)\mathcal{N}_{rs}(t), \tag{26}$$

where $\mathcal{N}_{rs}(t)$ is the additive noise indexed by the receivers and the sources. The latter index reminds us that the noise varies from one illumination to another. We suppose as before that $\mathcal{N}_{rs}(t)$ is white noise, independent over the array sensors

$$\mathbb{E}\left[\mathcal{N}_{rs}(t)\right] = 0, \qquad \mathbb{E}\left[\mathcal{N}_{rs}(t)\mathcal{N}_{r's'}(t')\right] = \sigma_{\mathcal{N}}^2\delta(t-t')\delta_{r,r'}\delta_{s,s'}. \tag{27}$$

We also choose a large enough T, so that we can neglect the effect of the window in the first term of the right-hand side of (26).

The forward map $\mathcal{M} : L^2(I_W) \to \mathbb{D}$ takes square integrable reflectivity functions ρ to functions in the data space

$$\mathbb{D} = \left\{ \hat{d}(\omega, \vec{\mathbf{x}}_r, \vec{\mathbf{x}}_s) \in \mathbb{C}, \quad \sum_{r=1}^{N_r} \sum_{s=1}^{N_s} \int_{|\omega-\omega_o|\le \pi B} d\omega \left| \hat{d}(\omega, \vec{\mathbf{x}}_r, \vec{\mathbf{x}}_s) \right|^2 < \infty \right\}, \quad (28)$$

with complex inner product

$$\left\langle \hat{d}, \hat{g} \right\rangle = \int_{|\omega-\omega_o|\le \pi B} d\omega \sum_{r=1}^{N_r} \sum_{s=1}^{N_s} \overline{\hat{d}(\omega, \vec{\mathbf{x}}_r, \vec{\mathbf{x}}_s)} \hat{g}(\omega, \vec{\mathbf{x}}_r, \vec{\mathbf{x}}_s), \quad \forall\ \hat{d}, \hat{g} \in \mathbb{D}, \quad (29)$$

and induced norm $\|\hat{d}\|_{\mathbb{D}} = \sqrt{\left\langle \hat{d}, \hat{d} \right\rangle}$. It is given by

$$[\mathcal{M}\rho](\omega, \vec{\mathbf{x}}_r, \vec{\mathbf{x}}_s) = k^2 \hat{f}(\omega) \int_{I_W} d\vec{\mathbf{y}} \rho(\vec{\mathbf{y}}) \hat{G}(\omega, \vec{\mathbf{x}}_r, \vec{\mathbf{y}}) \hat{G}(\omega, \vec{\mathbf{y}}, \vec{\mathbf{x}}_s), \quad (30)$$

and it assumes that we know both the pulse $\hat{f}(\omega)$ and the Green's function, i.e., $c(\vec{\mathbf{x}})$.

Least Squares Inversion

The imaging problem is to estimate from the array measurements the location of the unknown remote sources and/or reflectors in the medium. We show here how it relates to the problem of least squares minimization of the misfit between the measured data and the model prediction. Because the data space depends on the type of the array, we treat separately passive and active arrays. In both cases the unknown lies in the Hilbert space $L^2(I_W)$ with inner product (17).

Connection to Bayesian Inversion

The least squares minimization of the misfit between the measured data and the model prediction is widely used in the literature. It arises in Bayesian inversion, where the least squares minimizer is known as the *maximum likelihood* estimator.

Denote by $\pi(d\,|\rho)$ the density of the *likelihood function*, which is the probability that we observe the data d given ρ. Bayes' rule says that the *posterior probability* density $\pi(\rho|d)$ of ρ, given the data, is

$$\pi(\rho|d) = \frac{\pi(d\,|\rho)\pi(\rho)}{\pi(d)},$$

where $\pi(\rho)$ is the prior probability density of ρ and $\pi(d)$ is the marginal likelihood density of the data. Assuming no information about ρ prior to the measurements, we take a uniform density $\pi(\rho)$ over the imaging region. The marginal likelihood $\pi(d)$ tells us if the data can be achieved by measurements. We suppose that this is so, which means that $\pi(d)$ is just a constant which plays no role in the estimation.

Under all these assumptions, we obtain that maximizing the posterior $\pi(\rho|d)$ is the same as maximizing the likelihood function $\pi(d|\rho)$. This is given by

$$\pi(d|\rho) \sim \exp\left[-\frac{\|d - \mathcal{M}\rho\|_{\mathbb{D}}^2}{2\sigma_T^2}\right], \tag{31}$$

and its maximizer, the maximum likelihood estimate, is the same as the least squares solution described in the next sections. The symbol $\sim$ in Eq. (31) means equal up to a normalization constant. We have a normal probability density because we assume white noise which is Gaussian and independent over the receivers. The variance $\sigma_T^2 = T\sigma_{\mathcal{N}}^2$ is the noise power at a sensor, calculated over the duration T of the measurements. The noise is essentially uncorrelated over the frequencies when the bandwidth is sampled at intervals $\Delta\omega \gg 2\pi/T$, as follows by a straightforward calculation.

Imaging with Passive Arrays

The variational formulation of the inverse problem is: Find $\rho \in L^2(I_W)$, a minimizer of the data misfit

$$\mathcal{O}(\rho) = \|\hat{d} - \mathcal{M}\rho\|_{\mathbb{D}}^2 = \int_{|\omega-\omega_o|\le\pi B} d\omega \sum_{r=1}^{N_r} \left|\hat{d}(\omega, \vec{\mathbf{x}}_r) - [\mathcal{M}\rho](\omega, \vec{\mathbf{x}}_r)\right|^2. \tag{32}$$

We may add a penalty term in (32), such as a multiple of $\|\rho\|_2^2$ to regularize the problem. We do not do so because it does not play a role in the basic imaging methods derived below. However, regularization is important and one can use other penalties, such as total variation and sparsity promoting norms, to improve the quality of images [15,26].

The first-order optimality condition satisfied by the minimizer of (32) is

$$[\mathcal{M}^\star\mathcal{M}\rho](\vec{\mathbf{y}}^s) = \left[\mathcal{M}^\star\hat{d}\right](\vec{\mathbf{y}}^s), \quad \forall\, \vec{\mathbf{y}}^s \in I_W, \tag{33}$$

where the operator $\mathcal{M}^\star : \mathbb{D} \to L^2(I_W)$ is the adjoint of $\mathcal{M}$, defined formally by the relation

$$\left\langle \hat{d}\,, \mathcal{M}\rho \right\rangle = \left(\mathcal{M}^\star\hat{d}\,, \rho\right), \quad \forall\, \rho \in L^2(I_W), \;\; \hat{d} \in \mathbb{D}. \tag{34}$$

It is given explicitly by

$$\left[\mathcal{M}^\star\hat{d}\right](\vec{\mathbf{y}}^s) = \int_{|\omega-\omega_o|\le\pi B} d\omega\, \hat{f}(\omega) \sum_{r=1}^{N_r} \overline{\hat{d}(\omega, \vec{\mathbf{x}}_r)}\hat{G}(\omega, \vec{\mathbf{y}}^s, \vec{\mathbf{x}}_r), \tag{35}$$

where $\vec{\mathbf{y}}^s$ are the *search points* that sweep the imaging region I_W, and the bar denotes the complex conjugate.

Equation (33) are the normal equations and $\mathcal{M}^\star\mathcal{M} : L^2(I_W) \to L^2(I_W)$ is the *normal operator*

$$[\mathcal{M}^\star\mathcal{M}\rho](\vec{\mathbf{y}}^s) = \int_{I_W} d\vec{\mathbf{y}}\, \rho(\vec{\mathbf{y}}) \mathcal{K}_{\rm pas}(\vec{\mathbf{y}}, \vec{\mathbf{y}}^s), \tag{36}$$

with kernel

$$\mathcal{K}_{\rm pas}(\vec{\mathbf{y}}, \vec{\mathbf{y}}^s) = \int_{|\omega-\omega_o|\le \pi B} d\omega \left|\hat{f}(\omega)\right|^2 \sum_{r=1}^{N_r} \overline{\hat{G}(\omega, \vec{\mathbf{y}}, \vec{\mathbf{x}}_r)} \hat{G}(\omega, \vec{\mathbf{y}}^s, \vec{\mathbf{x}}_r). \tag{37}$$

It is usually not invertible, so one way of solving (33) is to compute the pseudo-inverse of $\mathcal{M}^\star\mathcal{M}$, which is typically very expensive. However, the kernel $\mathcal{K}_{\rm pas}(\vec{\mathbf{y}}, \vec{\mathbf{y}}^s)$ has the important property that is large for points $\vec{\mathbf{y}}$ in the vicinity of $\vec{\mathbf{y}}^s$ and small otherwise, as we explain in more detail in section "The Normal Operator and the Time Reversal Process." Thus, from the point of view of approximating the support of ρ, we may replace the normal operator in (33) by the identity, and obtain that

$$\rho(\vec{\mathbf{y}}^s) \sim \left[\mathcal{M}^\star \hat{d}\,\right](\vec{\mathbf{y}}^s) = \int_{|\omega-\omega_o|\le \pi B} d\omega\, \hat{f}(\omega) \sum_{r=1}^{N_r} \overline{\hat{d}(\omega, \vec{\mathbf{x}}_r)} \hat{G}(\omega, \vec{\mathbf{y}}^s, \vec{\mathbf{x}}_r). \tag{38}$$

The factor $\hat{f}(\omega)$ is not essential in imaging with pulses, although it is important when the source emits a long chirped signal or a stationary noise signal. In that case the factor $\hat{f}(\omega)$ arises in a data pre-processing step known as pulse compression. Since we assume that the source emits a pulse, we may neglect $\hat{f}(\omega)$ in (38) to obtain

$$\rho(\vec{\mathbf{y}}^s) \sim \mathcal{J}(\vec{\mathbf{y}}^s) = \int_{|\omega-\omega_o|\le \pi B} \frac{d\omega}{2\pi} \sum_{r=1}^{N_r} \overline{\hat{d}(\omega, \vec{\mathbf{x}}_r)} \hat{G}(\omega, \vec{\mathbf{y}}^s, \vec{\mathbf{x}}_r). \tag{39}$$

This is useful for applications where $f(t)$ may not be known. The normalization constant $1/(2\pi)$ is convenient for inverting the Fourier transform.

The meaning of the symbol $\sim$ in (38) and (39) is that large values of the right-hand sides correspond to points in the vicinity of the support of ρ. Thus, $\left[\mathcal{M}^\star \hat{d}\,\right](\vec{\mathbf{y}}^s)$ or its simplification $\mathcal{J}(\vec{\mathbf{y}}^s)$ are imaging functions.

Imaging with Active Arrays

The case of active arrays is similar to the above, except that the data space is given by (28), and the complex inner product is (29). The unknown reflectivity ρ is estimated by a minimizer of

$$\mathcal{O}(\rho) = \|\hat{d} - \mathcal{M}\rho\|_{\mathbb{D}}^2 = \int_{|\omega-\omega_o|\le\pi B} d\omega \sum_{r=1}^{N_r}\sum_{s=1}^{N_s} \left|\hat{d}\,(\omega,\vec{\mathbf{x}}_r,\vec{\mathbf{x}}_s) - [\mathcal{M}\rho]\,(\omega,\vec{\mathbf{x}}_r,\vec{\mathbf{x}}_s)\right|^2, \tag{40}$$

and satisfies the normal equations

$$[\mathcal{M}^\star\mathcal{M}\rho]\,(\vec{\mathbf{y}}^s) = \left[\mathcal{M}^\star\hat{d}\,\right](\vec{\mathbf{y}}^s), \quad \forall\ \vec{\mathbf{y}}^s \in I_W. \tag{41}$$

The adjoint operator $\mathcal{M}^\star : \mathbb{D} \to L^2(I_W)$ defined formally by

$$\left\langle \hat{d}\,, \mathcal{M}\rho \right\rangle = \left(\mathcal{M}^\star\hat{d}\,, \rho\right), \quad \forall\ \rho \in L^2(I_W),\ \ \hat{d} \in \mathbb{D}, \tag{42}$$

has the explicit expression

$$\left[\mathcal{M}^\star\hat{d}\,\right](\vec{\mathbf{y}}^s) = \int_{|\omega-\omega_o|\le\pi B} d\omega\, k^2\,\hat{f}(\omega) \sum_{r=1}^{N_r}\sum_{s=1}^{N_s} \overline{\hat{d}\,(\omega,\vec{\mathbf{x}}_r,\vec{\mathbf{x}}_s)}\hat{G}(\omega,\vec{\mathbf{y}}^s,\vec{\mathbf{x}}_r) \tag{43}$$
$$\hat{G}(\omega,\vec{\mathbf{y}}^s,\vec{\mathbf{x}}_s),$$

and the normal operator $\mathcal{M}^\star\mathcal{M} : L^2(I_W) \to L^2(I_W)$,

$$[\mathcal{M}^\star\mathcal{M}\rho]\,(\vec{\mathbf{y}}^s) = \int_{I_W} d\vec{\mathbf{y}}\,\rho(\vec{\mathbf{y}})\mathcal{K}_{\text{act}}(\vec{\mathbf{y}},\vec{\mathbf{y}}^s), \tag{44}$$

has the kernel

$$\mathcal{K}_{\text{act}}(\vec{\mathbf{y}},\vec{\mathbf{y}}^s) = \int_{|\omega-\omega_o|\le\pi B} d\omega\, k^4 \left|\hat{f}(\omega)\right|^2 \sum_{r=1}^{N_r} \overline{\hat{G}(\omega,\vec{\mathbf{y}},\vec{\mathbf{x}}_r)}\hat{G}(\omega,\vec{\mathbf{y}}^s,\vec{\mathbf{x}}_r)$$
$$\sum_{s=1}^{N_s} \overline{\hat{G}(\omega,\vec{\mathbf{y}},\vec{\mathbf{x}}_s)}\hat{G}(\omega,\vec{\mathbf{y}}^s,\vec{\mathbf{x}}_s). \tag{45}$$

We discuss this kernel in some detail in section "The Normal Operator and the Time Reversal Process," and we relate it to the kernel (37) arising in imaging with passive arrays. Again, we expect that $\mathcal{K}_{\text{act}}$ peaks near its diagonal, so we can estimate the support of the reflectivity by replacing $\mathcal{M}^\star\mathcal{M}$ with the identity in the normal equations. We obtain

$$\rho(\vec{\mathbf{y}}^s) \sim \left[\mathcal{M}^\star\hat{d}\,\right](\vec{\mathbf{y}}^s) = \int_{|\omega-\omega_o|\le\pi B} d\omega\, k^2\,\hat{f}(\omega) \sum_{r=1}^{N_r}\sum_{s=1}^{N_s} \overline{\hat{d}\,(\omega,\vec{\mathbf{x}}_r,\vec{\mathbf{x}}_s)}\hat{G}(\omega,\vec{\mathbf{y}}^s,\vec{\mathbf{x}}_r)$$
$$\hat{G}(\omega,\vec{\mathbf{y}}^s,\vec{\mathbf{x}}_s), \tag{46}$$

with symbol $\sim$ having the same meaning as in (39).

The factor $\omega^2 \hat{f}(\omega)$ in (46) amounts to convolving the second derivative $f''(t)$ of the emitted signal with the time reversed data. But in the Born approximation (26) the time reversed data is determined by $f''(-t)$ convolved with the Green's functions. Thus, we have $f''(t) \star_t f''(-t)$ in (46). Such convolutions are important only when the emitted signals are long, like chirps, or noise signals. Because our $f(t)$ is a pulse, we can estimate the support of the reflectivity using the imaging function

$$\rho(\vec{\mathbf{y}}^s) \sim \mathcal{J}(\vec{\mathbf{y}}^s) = \int_{|\omega-\omega_o|\le \pi B} \frac{d\omega}{2\pi} \sum_{r=1}^{N_r} \sum_{s=1}^{N_s} \overline{\hat{d}(\omega, \vec{\mathbf{x}}_r, \vec{\mathbf{x}}_s)} \hat{G}(\omega, \vec{\mathbf{y}}^s, \vec{\mathbf{x}}_r) \hat{G}(\omega, \vec{\mathbf{y}}^s, \vec{\mathbf{x}}_s). \tag{47}$$

The Normal Operator and the Time Reversal Process

We draw here an analogy between the normal operator and the time reversal process, to give a physical meaning of its replacement by the identity in Eqs. (38) and (46). The time reversal process is described in section "The Time Reversal Process." The connection to the normal operator is in section "The Normal Operator."

The Time Reversal Process

The time reversal process is a physical experiment that uses an array of N sensors which act as both receivers and sources. It is a two-step process. In the first step the sensors are receivers which record over a time window $\chi_T(t)$ of duration T, the waves generated by a remote point-like source at $\vec{\mathbf{y}}$ that emits a pulse $\varphi(t)$. This may be the same as $f(t)$, but it is convenient for the discussion to take an arbitrary function $\varphi(t)$, with the same bandwidth as $f(t)$. We can model the density of the point-like source by letting ρ be supported in a ball $B_R(\vec{\mathbf{y}})$ centered at $\vec{\mathbf{y}}$ of radius $R \ll \lambda_o$. This allows us to approximate the Green's function in the model of the data by its value at $\vec{\mathbf{y}}$, and obtain a net source amplitude equal to the integral of ρ over $B_R(\vec{\mathbf{y}})$. The wave equation and the imaging function are linear in ρ, so we normalize the net source amplitude to one.

In the second step of the experiment, the array time reverses the recordings and re-emits them simultaneously from all the sensors, which act as sources. The resulting wave propagates in the medium, and it refocuses near the original source, due to the time reversibility of the wave equation. The quality of the refocusing depends on how large the array is, the length of time of the recordings, and the medium.

The recordings at the receivers are denoted by

$$d_\chi(t, \vec{\mathbf{x}}_r) = \chi_T(t)\varphi(t) \star_t G(t, \vec{\mathbf{x}}_r, \vec{\mathbf{y}}), \tag{48}$$

and their time reversed version $d_{TR}(t,\vec{\mathbf{x}}_r) = d_\chi(T-t,\vec{\mathbf{x}}_r)$ has the Fourier transform

$$\hat{d}_{TR}(\omega,\vec{\mathbf{x}}_r) = e^{i\omega T}\overline{\hat{d}_\chi(\omega,\vec{\mathbf{x}}_r)}, \tag{49}$$

for $r = 1,\ldots,N$. This is the ideal model, without additive noise. We obtain from (48) that

$$\hat{d}_\chi(\omega,\vec{\mathbf{x}}_r) = \int \frac{d\omega'}{2\pi}\hat{\chi}(\omega' T)\hat{\varphi}(\omega-\omega')\hat{G}(\omega-\omega',\vec{\mathbf{x}}_r,\vec{\mathbf{y}}) \approx \hat{\varphi}(\omega)\hat{G}(\omega,\vec{\mathbf{x}}_r,\vec{\mathbf{y}})\,, \tag{50}$$

where we let

$$\chi_T(t) = \frac{1}{T}\chi\left(\frac{t}{T}\right),$$

for a function χ of dimensionless argument and support of order one. The approximation in (50) is for large recording times T satisfying $\omega_o T \gg 1$, and for the window normalization $\chi_T(0) = 1$.

The array emits the time reversed field d_{TR} from all the sensors at once, and the wave observed at the search points $\vec{\mathbf{y}}^s$ is modeled by

$$\begin{aligned} p_{TR}(T+t,\vec{\mathbf{y}}^s,\vec{\mathbf{y}}) &= \sum_{r=1}^N \int \frac{d\omega}{2\pi}\hat{d}_{TR}(\omega,\vec{\mathbf{x}}_r)\hat{G}(\omega,\vec{\mathbf{x}}_r,\vec{\mathbf{y}}^s)e^{-i\omega(T+t)} \\ &= \sum_{r=1}^N \int \frac{d\omega}{2\pi}\overline{\hat{d}_\chi(\omega,\vec{\mathbf{x}}_r)}\hat{G}(\omega,\vec{\mathbf{x}}_r,\vec{\mathbf{y}}^s)e^{-i\omega t} \\ &\approx \sum_{r=1}^N \int \frac{d\omega}{2\pi}\overline{\hat{\varphi}(\omega)}\,\overline{\hat{G}(\omega,\vec{\mathbf{x}}_r,\vec{\mathbf{y}})}\hat{G}(\omega,\vec{\mathbf{x}}_r,\vec{\mathbf{y}}^s)e^{-i\omega t}, \end{aligned} \tag{51}$$

where we used the approximation (50). Because of the time reversibility of the wave equation, it refocuses at the original source, at offset time $t = 0$. The *time reversal point spread function*

$$\mathcal{K}_{TR}(\vec{\mathbf{y}}^s,\vec{\mathbf{y}}) = p_{TR}(T,\vec{\mathbf{y}}^s,\vec{\mathbf{y}}) \approx \sum_{r=1}^N \int \frac{d\omega}{2\pi}\overline{\hat{\varphi}(\omega)}\,\overline{\hat{G}(\omega,\vec{\mathbf{x}}_r,\vec{\mathbf{y}})}\hat{G}(\omega,\vec{\mathbf{x}}_r,\vec{\mathbf{y}}^s) \tag{52}$$

models the refocused wave field. We describe it briefly in section "Imaging in Smooth and Known Media" in the case of smooth media. The cross-range resolution of the refocusing depends on the wavelength, the distance from the source to the array and the array aperture. The range resolution is determined by the bandwidth. In complex media the focusing may be improved. This is known as *super-resolution*.

By improved we mean that with a limited aperture we may get better cross-range resolution than we would if the medium were smooth. We refer to [6, 19, 27] for the analysis and illustration of super-resolution.

The Normal Operator

In the case of passive arrays, the normal operator is given by (36). It is a linear integral operator with kernel $\mathcal{K}_{\rm pass}(\vec{\mathbf{y}}, \vec{\mathbf{y}}^s)$ written explicitly in (37). We see that it is the same as the time reversal point spread function (52) for $\varphi(\omega) = |\hat{f}(\omega)|^2$. That is to say, the kernel of the normal operator is mathematically equivalent to the time reversal function for a source at $\vec{\mathbf{y}}$, emitting the pulse with Fourier transform equal to $|\hat{f}(\omega)|^2$. The tighter the refocusing of the wave at $\vec{\mathbf{y}}$ in the time reversal process, the more the kernel is peaked at $\vec{\mathbf{y}} = \vec{\mathbf{y}}^s$, and the closer the behavior of the normal operator to an approximate identity in Eq. (38).

The normal operator for the case of active arrays is given by (44), and its kernel is (45). Let us write it assuming that each sensor in the array acts as a source and receiver, so that $N_r = N_s = N$. We have

$$\mathcal{K}_{\rm act}(\vec{\mathbf{y}}, \vec{\mathbf{y}}^s) = \int_{|\omega-\omega_o|\le \pi B} d\omega\, k^4 |\hat{f}(\omega)|^2 \left[\sum_{r=1}^N \overline{\hat{G}(\omega, \vec{\mathbf{x}}_r, \vec{\mathbf{y}})} \hat{G}(\omega, \vec{\mathbf{y}}^s, \vec{\mathbf{x}}_r)\right]^2, \tag{53}$$

which can be related to the time reversal process using the time-dependent field

$$p_{TR}^{(N)}(T+t, \vec{\mathbf{y}}, \vec{\mathbf{y}}^s) = \int_{|\omega-\omega_o|\le \pi B} d\omega\, k^2 |\hat{f}(\omega)| \sum_{r=1}^N \overline{\hat{G}(\omega, \vec{\mathbf{x}}_r, \vec{\mathbf{y}})} \hat{G}(\omega, \vec{\mathbf{y}}^s, \vec{\mathbf{x}}_r) e^{-i\omega t}. \tag{54}$$

This is the wave observed at location $\vec{\mathbf{y}}^s$ and time $T+t$ in the time reversal process, in the case of a source at $\vec{\mathbf{y}}$, emitting a pulse with Fourier coefficient $k^2|\hat{f}(\omega)|$. The focusing in (54) is expected at $t = 0$, and the time reversal point spread function is equal to $p_{TR}^{(N)}(T, \vec{\mathbf{y}}, \vec{\mathbf{y}}^s)$. The kernel (53) is, up to a constant, the time convolution of $p_{TR}^{(N)}$ with itself, evaluated at $t = 0$,

$$\mathcal{K}_{\rm act}(\vec{\mathbf{y}}, \vec{\mathbf{y}}^s) = 2\pi p_{TR}^{(N)}(T+t, \vec{\mathbf{y}}, \vec{\mathbf{y}}^s) \star_t p_{TR}^{(N)}(T+t, \vec{\mathbf{y}}, \vec{\mathbf{y}}^s)\big|_{t=0}. \tag{55}$$

If the wave field $p_{TR}^{(N)}(T+t, \vec{\mathbf{y}}, \vec{\mathbf{y}}^s)$ focuses at $\vec{\mathbf{y}}^s = \vec{\mathbf{y}}$ around $t = 0$, so will the kernel. Thus, the normal operator behaves as an approximate identity if the time reversal process focuses with sharp resolution.

The generalization of (55) to arrays with N_r receivers and N_s sources is

$$\mathcal{K}_{\rm act}(\vec{\mathbf{y}}, \vec{\mathbf{y}}^s) = 2\pi p_{TR}^{(N_r)}(T+t, \vec{\mathbf{y}}, \vec{\mathbf{y}}^s) \star_t p_{TR}^{(N_s)}(T+t, \vec{\mathbf{y}}, \vec{\mathbf{y}}^s)\big|_{t=0}. \tag{56}$$

In particular, for a single source $N_s = 1$, and there is no refocusing of $p_{TR}^{(N_s)}(T + t, \vec{\mathbf{y}}, \vec{\mathbf{y}}^s)$ at $\vec{\mathbf{y}}^s = \vec{\mathbf{y}}$. The behavior of the kernel is determined by the refocusing of the wave $p_{TR}^{(N_r)}(T + t, \vec{\mathbf{y}}, \vec{\mathbf{y}}^s)$.

Imaging in Smooth and Known Media

In known media the imaging function (39) is the mathematical analogue of the time reversal point spread function. We describe it in this section for the case of smooth media. In unknown media we cannot compute (39). We use instead the *reverse time migration* imaging function that propagates the waves from the array to the imaging region in a surrogate medium. It is the method of choice in many imaging applications. The *Kirchhoff migration* imaging function is a high frequency approximation of the reverse time migration function, where the propagation is replaced by a synchronization of the waves using travel times.

Once we have defined the imaging functions, the question is how well do they work? To answer it, we identify two properties that make an imaging method useful:

1. The image should focus at ρ, i.e., $|\mathcal{J}(\vec{\mathbf{y}}^s)|$ should be large near the support of ρ and small elsewhere.
2. The focusing should be robust.

We study in section "Robustness to Additive Noise." We postpone the more involved discussion of *robustness to uncertainties in complex media* until Sects. 7 and 8. The characterization of the focusing of the image is known as *resolution analysis*. We state the resolution limits for a setup with small arrays.

Passive Arrays

When we know the medium, and therefore the Green's function $\hat{G}$, we can process the data as in (39) to form an image. The data processing has the physical interpretation of taking the traces received at the array, time reversing them, and then back-propagating them to the imaging point via the Green's function. Because the Green's function models the actual wave propagation in the known medium, the back-propagation in $\mathcal{J}(\vec{\mathbf{y}}^s)$ is equivalent to solving the wave equation with source term

$$F(t, \vec{\mathbf{x}}) = \sum_{r=1}^{N_r} d(-t, \vec{\mathbf{x}}_r)\delta(\vec{\mathbf{x}} - \vec{\mathbf{x}}_r),$$

and evaluating the pressure field at $\vec{\mathbf{y}}^s \in I_W$, at the expected time of refocus. The imaging process is mathematically equivalent to the time reversal process described in section "The Normal Operator and the Time Reversal Process" because we know precisely the medium!

To study the imaging function (39), let us recall the model (16) of the data and write

$$\mathcal{J}(\vec{\mathbf{y}}^s) = \int_{|\omega-\omega_o|\le\pi B} d\omega \sum_{r=1}^{N_r} \overline{\hat{d}(\omega,\vec{\mathbf{x}}_r)}\hat{G}(\omega,\vec{\mathbf{y}}^s,\vec{\mathbf{x}}_r) = \left\langle \mathcal{J}(\vec{\mathbf{y}}^s)\right\rangle + \delta\mathcal{J}(\vec{\mathbf{y}}^s). \tag{57}$$

The first term is the model of the image in the absence of noise

$$\left\langle \mathcal{J}(\vec{\mathbf{y}}^s)\right\rangle = \int_{I_W} d\vec{\mathbf{y}}\,\rho(\vec{\mathbf{y}}) \int_{|\omega-\omega_o|\le\pi B} d\omega\, \overline{\hat{f}(\omega)} \sum_{r=1}^{N_r} \overline{\hat{G}(\omega,\vec{\mathbf{x}}_r,\vec{\mathbf{y}})}\hat{G}(\omega,\vec{\mathbf{y}}^s,\vec{\mathbf{x}}_r), \tag{58}$$

and the second term models the effect of the noise

$$\delta\mathcal{J}(\vec{\mathbf{y}}^s) = \int_{|\omega-\omega_o|\le\pi B} d\omega \sum_{r=1}^{N_r} \overline{\hat{\mathcal{N}}_r^T}(\omega)\hat{G}(\omega,\vec{\mathbf{y}}^s,\vec{\mathbf{x}}_r). \tag{59}$$

We study $\delta\mathcal{J}(\vec{\mathbf{y}}^s)$ in section "Robustness to Additive Noise," so we concentrate here on $\left\langle \mathcal{J}(\vec{\mathbf{y}}^s)\right\rangle$, which we rewrite as

$$\left\langle \mathcal{J}(\vec{\mathbf{y}}^s)\right\rangle \approx \int_{I_W} d\vec{\mathbf{y}}\rho(\vec{\mathbf{y}}) \int_{|\omega-\omega_o|\le\pi B} \frac{d\omega}{2\pi}\, \overline{\hat{f}(\omega)} \sum_{r=1}^{N_r} \overline{\alpha(\vec{\mathbf{x}}_r,\vec{\mathbf{y}})}\alpha(\vec{\mathbf{x}}_r,\vec{\mathbf{y}}^s) e^{i\omega[\tau(\vec{\mathbf{x}}_r,\vec{\mathbf{y}}^s)-\tau(\vec{\mathbf{x}}_r,\vec{\mathbf{y}})]} \tag{60}$$

$$= \int_{I_W} d\vec{\mathbf{y}}\rho(\vec{\mathbf{y}}) \sum_{r=1}^{N_r} \overline{\alpha(\vec{\mathbf{x}}_r,\vec{\mathbf{y}})}\alpha(\vec{\mathbf{x}}_r,\vec{\mathbf{y}}^s) f\left(\tau(\vec{\mathbf{x}}_r,\vec{\mathbf{y}}) - \tau(\vec{\mathbf{x}}_r,\vec{\mathbf{y}}^s)\right). \tag{61}$$

Here we used the geometrical optics approximation of the Green's function in the assumed smooth medium

$$\hat{G}(\omega,\vec{\mathbf{x}},\vec{\mathbf{y}}) \approx \alpha(\vec{\mathbf{x}},\vec{\mathbf{y}}) e^{i\omega\tau(\vec{\mathbf{x}},\vec{\mathbf{y}})}, \tag{62}$$

with amplitude α and travel time τ. They are calculated along rays (geodesics) connecting $\vec{\mathbf{x}}$ to $\vec{\mathbf{y}}$ by solving with the method of characteristics a transport equation and an eikonal equation [5,24]. In the case of homogeneous media with wave speed c_o, they have the explicit expression

$$\alpha(\vec{\mathbf{x}},\vec{\mathbf{y}}) = \alpha_o(\vec{\mathbf{x}},\vec{\mathbf{y}}) := \frac{1}{4\pi|\vec{\mathbf{x}}-\vec{\mathbf{y}}|} \quad \text{and} \quad \tau(\vec{\mathbf{x}},\vec{\mathbf{y}}) = \tau_o(\vec{\mathbf{x}},\vec{\mathbf{y}}) := \frac{|\vec{\mathbf{x}}-\vec{\mathbf{y}}|}{c_o}. \tag{63}$$

We can understand what to expect from (61) by referring to the example in Fig. 1, with a source of small support around point $\vec{\mathbf{y}}$. Because data $d(t,\vec{\mathbf{x}}_r)$ are essentially the emitted pulse delayed by the travel time $\tau(\vec{\mathbf{x}}_r,\vec{\mathbf{y}})$, we get a contribution to the

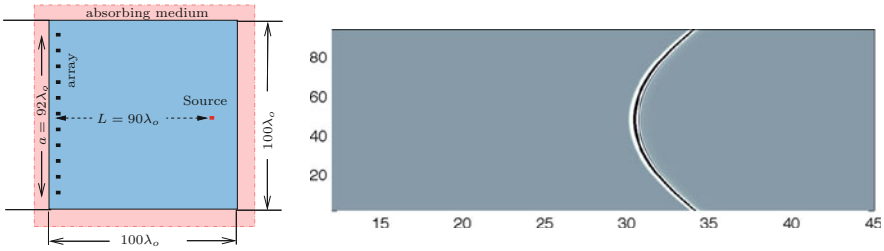


Fig. 1 Plot of the passive array data traces for a point source in a homogeneous medium. They solve the wave equation in a homogeneous medium with wave speed $c_o = 1.5$ km/s, for a source emitting a pulse with wide frequency band [1.5–4.5] MHz. We plot $d(t, \vec{\mathbf{x}}_r)$ in *grey scale*, as a function of time in the abscissa and receiver location in the ordinate. All the lengths are scaled by the central wavelength $\lambda_o = 0.5$ mm. The simulation setup is shown on the *left*

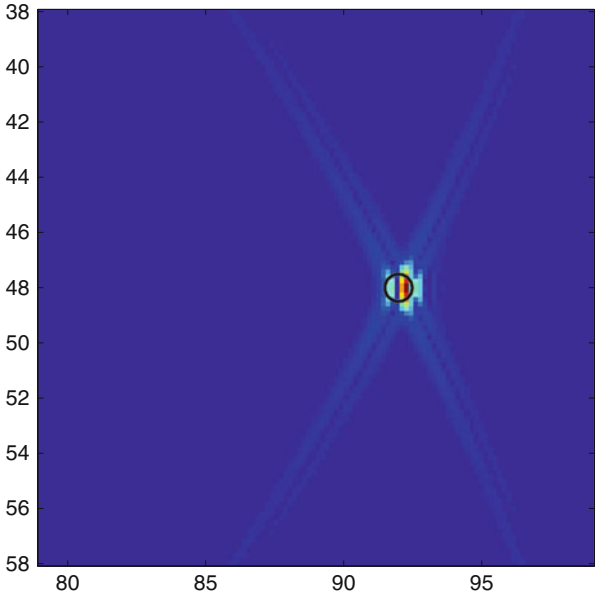


Fig. 2 Kirchhoff migration image for one point source in a homogeneous medium. This is the image obtained using (64) for the passive array data traces shown in Fig. 1. We plot the absolute value of the imaging function. The abscissa is range in λ_o and the ordinate is cross-range in λ_o

sum in (61) if $\vec{\mathbf{y}}^s$ lies near the isochrone with travel time equal to $\tau(\vec{\mathbf{x}}_r, \vec{\mathbf{y}})$, at least for one receiver. For example, in homogeneous media this means that $\vec{\mathbf{y}}^s$ is near the sphere of center $\vec{\mathbf{x}}_r$ and radius $|\vec{\mathbf{x}}_r - \vec{\mathbf{y}}|$, at least for one receiver. With this reasoning, we expect that $\left|\mathcal{J}(\vec{\mathbf{y}}^s)\right|$ is large near the intersection of all these isochrones, for all receivers in the array. Such points are close to $\vec{\mathbf{y}}$, as we illustrate in Fig. 2. The imaging function for point-like sources is called the *point spread function* and its essential support around $\vec{\mathbf{y}}$ defines the resolution limits.

We assume in this article a typical imaging regime with an array of linear size a that is much smaller than the distance to the source, and with bandwidth $B < \omega_o$. The direction from the source to the array is called *range*, and the distance along it

is of the order of the range scale L. The directions orthogonal to the range are called *cross-range*. The support of the point spread function in cross-range, the *cross-range resolution*, is of the order $\lambda_o L/a$. It improves as we increase the frequency and the array aperture. The range resolution depends on the temporal support of the pulse, which defines the precision of the arrival time estimation of the waves, and thus of the distance to the source. Because the pulse $f(t)$ has the temporal support $O(1/B)$, it is of the order c_o/B. These resolution limits are widely known and can be deduced from the calculations in Sect. 7.

A common imaging approach in the applied literature is a further simplification of $\mathcal{J}^M(\vec{\mathbf{y}}^s)$, known as *Kirchhoff migration* [4],

$$\mathcal{J}^{\rm KM}(\vec{\mathbf{y}}^s) = \sum_{r=1}^{N_r} d\left(\tau(\vec{\mathbf{x}}_r, \vec{\mathbf{y}}^s), \vec{\mathbf{x}}_r\right). \tag{64}$$

It is similar to reverse time migration except that it neglects the geometrical spreading factor $\alpha(\vec{\mathbf{y}}^s, \vec{\mathbf{x}}_r)$ in the Green's function, as if it were a constant across the array. This approximation is justified in our setup because the array aperture is much smaller than the distance between the array and the imaging region.

Active Arrays

The *reverse time migration* function for active arrays is

$$\begin{aligned}
\mathcal{J}^M(\vec{\mathbf{y}}^s) &= \int_{|\omega-\omega_o|\le \pi B} \frac{d\omega}{2\pi} \sum_{r=1}^{N_r}\sum_{s=1}^{N_s} \overline{\hat{d}(\omega, \vec{\mathbf{x}}_r, \vec{\mathbf{x}}_s)}\hat{G}(\omega, \vec{\mathbf{x}}_r, \vec{\mathbf{y}}^s)\hat{G}(\omega, \vec{\mathbf{x}}_s, \vec{\mathbf{y}}^s) \\
&\approx \int_{|\omega-\omega_o|\le \pi B} \frac{d\omega}{2\pi} \sum_{r=1}^{N_r}\sum_{s=1}^{N_s} \overline{\hat{d}(\omega, \vec{\mathbf{x}}_r, \vec{\mathbf{x}}_s)}\alpha(\vec{\mathbf{x}}_r, \vec{\mathbf{y}}^s)\alpha(\vec{\mathbf{x}}_s, \vec{\mathbf{y}}^s) e^{i\omega\left[\tau(\vec{\mathbf{x}}_r, \vec{\mathbf{y}}^s)+\tau(\vec{\mathbf{x}}_s, \vec{\mathbf{y}}^s)\right]} \\
&= \sum_{r=1}^{N_r}\sum_{s=1}^{N_s} d(\tau(\vec{\mathbf{x}}_r, \vec{\mathbf{y}}^s) + \tau(\vec{\mathbf{x}}_s, \vec{\mathbf{y}}^s), \vec{\mathbf{x}}_r, \vec{\mathbf{x}}_s)\alpha(\vec{\mathbf{x}}_r, \vec{\mathbf{y}}^s)\alpha(\vec{\mathbf{x}}_s, \vec{\mathbf{y}}^s).
\end{aligned} \tag{65}$$

The *Kirchhoff migration* function neglects the geometrical spreading factors $\alpha(\vec{\mathbf{x}}_r, \vec{\mathbf{y}}^s)$ of the Green's function, and forms an image by simply summing the traces "synchronized" by the expected round trip travel time,

$$\mathcal{J}^{\rm KM}(\vec{\mathbf{y}}^s) = \sum_{r=1}^{N_r}\sum_{s=1}^{N_s} d(\tau(\vec{\mathbf{x}}_r, \vec{\mathbf{y}}^s) + \tau(\vec{\mathbf{x}}_s, \vec{\mathbf{y}}^s), \vec{\mathbf{x}}_r, \vec{\mathbf{x}}_s). \tag{66}$$

We show in Fig. 5 the Kirchhoff migration images obtained with the data traces in Figs. 3 and 4. In each case we indicate the location of the scatterers with a black circle (Fig. 5).

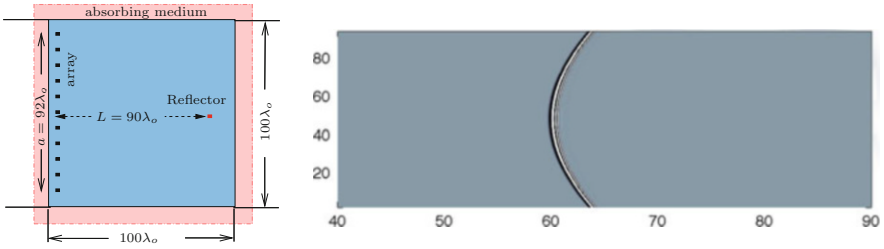


Fig. 3 Plot of the active array data traces for a point reflector in a homogeneous medium, and an illumination from the center element in $\mathcal{A}$. The abscissa is time in μs and the ordinate is the receiver location in λ_o. The simulation setup is shown on the *left*. It is the same as in Fig. 1 with the source being replaced by a soft acoustic reflector, i.e., a disk of diameter λ_o with homogeneous Dirichlet conditions on its boundary

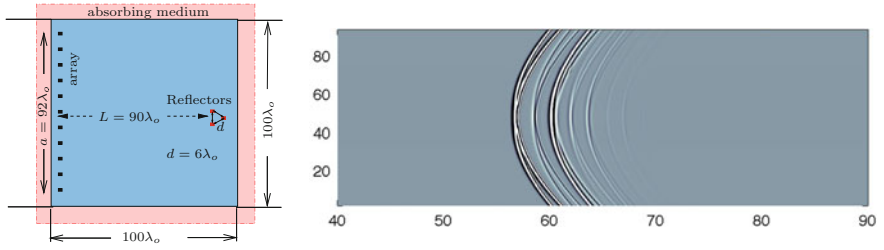


Fig. 4 Plot of the active array data traces for three reflectors in a homogeneous medium, and an illumination from the center element in the array. The abscissa is time in μs and the ordinate is the receiver location in λ_o. The simulation setup is shown on the *left*. The Born approximation used in the derivation of the imaging function captures only the primary echoes shown by the strong hyperbolae in the figure. The later arrivals are multiply scattered waves between the reflectors

We leave the interpretation of the imaging functions to the reader. They are straightforward extensions of the results for passive arrays.

Robustness to Additive Noise

To quantify the effect of the noise, we view the imaging problem in a stochastic framework, where $\mathcal{J}(\vec{\mathbf{y}}^s)$ is a random process due to the noise. The expectation of $\mathcal{J}(\vec{\mathbf{y}}^s)$ is its *coherent part* $\mathbb{E}\left[\mathcal{J}(\vec{\mathbf{y}}^s)\right] = \left\langle \mathcal{J}(\vec{\mathbf{y}}^s)\right\rangle$, and the random fluctuations are $\delta\mathcal{J}(\vec{\mathbf{y}}^s) = \mathcal{J}(\vec{\mathbf{y}}^s) - \mathbb{E}\left[\mathcal{J}(\vec{\mathbf{y}}^s)\right]$. Intuitively, robustness of the imaging method means that the coherent part $\langle\mathcal{J}\rangle$ dominates the fluctuations $\delta\mathcal{J}$ in the vicinity of its peaks, so we can determine with high fidelity the support of the unknown ρ. That is to say, the images do not change in essential ways with the realization of noise. In an imaging experiment we can compute only *one realization of the random process* $\mathcal{J}(\vec{\mathbf{y}}^s)$, which is what we call *an image*. It corresponds to the one realization of the noise in the data gathered by the receivers in the array. If we could compare such

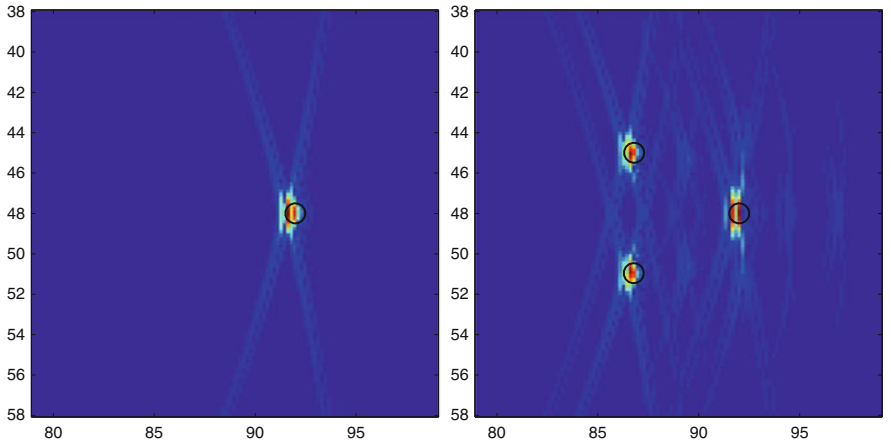


Fig. 5 Imaging results in homogeneous media using the traces plotted in Figs. 3 and 4. The images are computed with Eq. (66). The absolute value of the image for one reflector is shown on the *left* and for three reflectors on the *right*. The abscissa is range in λ_o and the ordinate is cross-range in λ_o. The imaging function is based on the Born approximation, and the image has some faint artifacts due to the multiply reflected waves seen in the right plot of Fig. 4

an image with $\langle \mathcal{J}(\vec{\mathbf{y}}^s) \rangle$, the ideal image with noiseless data, we would see many spurious peaks distributed throughout the search domain I_W, i.e., one realization of $\delta\mathcal{J}(\vec{\mathbf{y}}^s)$. If $|\delta\mathcal{J}|$ is smaller than $|\langle \mathcal{J} \rangle|$ for points in the support of ρ, with high probability, then the estimates of the support of ρ are essentially independent of the realization of the noise and the method is robust.

We assess below the robustness of images with passive arrays. The case of active arrays is similar. Recall Eqs. (57)–(59) and the assumptions (15) on $\mathcal{N}_r(t)$. Assume also a point-like source at $\vec{\mathbf{y}}$ in order to simplify the calculations. The coherent part of the image evaluated at $\vec{\mathbf{y}}$ is

$$\langle \mathcal{J}(\vec{\mathbf{y}}) \rangle \approx \int_{|\omega-\omega_o|\le \pi B} d\omega\, \overline{\hat{f}(\omega)} \sum_{r=1}^{N_r} \left| \overline{\hat{G}(\omega, \vec{\mathbf{x}}_r, \vec{\mathbf{y}})} \right|^2 \approx f(0) \sum_{r=1}^{N_r} |\alpha(\vec{\mathbf{x}}_r, \vec{\mathbf{y}})|^2, \tag{67}$$

and we wish that its magnitude be larger than $|\delta\mathcal{J}(\vec{\mathbf{y}})|$ with high probability. We can estimate this probability using Chebyshev's inequality

$$P\big[|\mathcal{J}(\vec{\mathbf{y}})| > \big|\langle \delta\mathcal{J}(\vec{\mathbf{y}}) \rangle\big|\big] = 1 - P\big[|\delta\mathcal{J}(\vec{\mathbf{y}})| \ge \big|\langle \mathcal{J}(\vec{\mathbf{y}}) \rangle\big|\big] \ge 1 - \frac{\mathbb{E}\big[|\delta\mathcal{J}(\vec{\mathbf{y}})|^2\big]}{\big|\langle \mathcal{J}(\vec{\mathbf{y}}) \rangle\big|^2} \tag{68}$$

$$= 1 - \left(\frac{1}{\mathrm{SNR}(\vec{\mathbf{y}})} \right)^2,$$

where SNR stands for the signal to noise ratio

$$\mathrm{SNR}(\vec{\mathbf{y}}) = \frac{\left|\left\langle \mathcal{J}(\vec{\mathbf{y}})\right\rangle\right|}{\sqrt{\mathrm{Var}\left[\mathcal{J}(\vec{\mathbf{y}})\right]}}, \qquad \mathrm{Var}\left[\mathcal{J}(\vec{\mathbf{y}})\right] = \sqrt{\mathbb{E}\left[|\delta\mathcal{J}(\vec{\mathbf{y}})|^2\right]}. \tag{69}$$

The "signal" is the coherent part (67) of the random process $\mathcal{J}(\vec{\mathbf{y}})$ and the "noise" is $\delta\mathcal{J}(\vec{\mathbf{y}})$. Its model follows from (59) and the geometrical optics approximation of the Green's function

$$\delta\mathcal{J}(\vec{\mathbf{y}}) \approx \sum_{r=1}^{N_r} \alpha(\vec{\mathbf{x}}_r, \vec{\mathbf{y}}^s) \int_{|\omega-\omega_o|\le\pi B} d\omega\, \overline{\hat{\mathcal{N}}_r^T(\omega)} e^{i\omega\tau(\vec{\mathbf{x}}_r,\vec{\mathbf{y}}^s)}. \tag{70}$$

It is a linear superposition of independent Gaussian noise terms, so it is Gaussian, with mean $\mathbb{E}\left[\delta\mathcal{J}(\vec{\mathbf{y}}^s)\right] = 0$ and variance

$$\begin{aligned}\mathrm{Var}\left[\mathcal{J}(\vec{\mathbf{y}}^s)\right] &\approx \sigma_{\mathcal{N}}^2 \sum_{r=1}^{N_r} |\alpha(\vec{\mathbf{x}}_r, \vec{\mathbf{y}}^s)|^2 \iint_{|\omega-\omega_o|\le\pi B} d\omega\, d\omega'\, e^{i(\omega-\omega')\tau(\vec{\mathbf{x}}_r,\vec{\mathbf{y}}^s)} \\ &\quad \int_0^T dt\, e^{-i(\omega-\omega')t} \\ &\approx 4\pi^2 B\sigma_{\mathcal{N}}^2 \sum_{r=1}^{N_r} |\alpha(\vec{\mathbf{x}}_r, \vec{\mathbf{y}}^s)|^2. \end{aligned} \tag{71}$$

The last approximation is for very large T and for $B\tau(\vec{\mathbf{x}}_r, \vec{\mathbf{y}}^s) \gg 1$, as is the case under our assumption that the temporal width $O(1/B)$ of the pulse is much smaller than the travel times from the source to the array.

The SNR follows from (67) and (71)

$$\mathrm{SNR}\left[\mathcal{J}(\vec{\mathbf{y}})\right] \approx \frac{|f(0)|}{2\pi\sigma_{\mathcal{N}}\sqrt{B}} \sqrt{\sum_{r=1}^{N} \left|\alpha(\vec{\mathbf{x}}_r, \vec{\mathbf{y}})\right|^2} \le \frac{\|f\|}{2\pi\sigma_{\mathcal{N}}} \sqrt{\sum_{r=1}^{N} \left|\alpha(\vec{\mathbf{x}}_r, \vec{\mathbf{y}})\right|^2}, \tag{72}$$

where

$$\|f\|^2 = \int dt\, |f(t)|^2 = \frac{1}{2\pi} \int_{|\omega-\omega_o|\le\pi B} d\omega\, |\hat{f}(\omega)|^2.$$

It increases linearly with the norm of the emitted signal and decreases with the "noise level" $\sigma_{\mathcal{N}}$. Having more sensors is beneficial because the sum over r increases. The amplitudes α decrease with the distance from the source to the array, so the noise level should be much lower than the norm of the emitted pulse in order to have high fidelity images of remote sources.

We neglect henceforth additive noise, for simplicity. Noise is of course part of any imaging experiment. It can be added easily in the analysis that follows and its effects can be studied as described above.

4 Challenges of Imaging in Complex Media

In the previous section we considered media that were largely known and smooth, with the exception of a compactly supported embedded reflector. In many imaging applications with active arrays the medium is complex, with wave speed $c(\vec{\mathbf{x}})$ that is not smooth and unknown. A natural idea is to estimate the function $c(\vec{\mathbf{x}})$ by the minimizer of the least squares functional $\left\| \mathcal{F}(c) - \hat{d} \right\|_{\mathbb{D}}^2$, where $\mathcal{F}$ is the *nonlinear forward mapping* that takes the wave speed c to the data space $\mathbb{D}$. Given the large number of unknowns, corresponding to a discretization of $c(\vec{\mathbf{x}})$ on a grid, the optimization would have to use a gradient-based optimization method like Gauss–Newton [23], which linearizes locally $\mathcal{F}(c)$. It turns out that it is easier to linearize $\mathcal{F}(c)$ with respect to localized perturbations of $c(\vec{\mathbf{x}})$ than with respect to distributed (smooth) ones. For example, the smooth part of $c(\vec{\mathbf{x}})$ determines the travel times τ of the waves, and in high frequency regimes small perturbations of τ give large changes of the wave field that oscillates like $e^{i\omega\tau}$. This indicates that it should be easier to estimate a sparse set of large discontinuities of $c(\vec{\mathbf{x}})$, which is the goal in imaging, than the rest of the function.

There is computational experience on the least squares minimization over $c(\vec{\mathbf{x}})$ in piecewise smooth media, and some convergence studies exist [30, 31, 34, 37]. The problem is difficult because the least squares functional is not convex, and the optimization gets stuck easily in local minima that have nothing to do with the true $c(\vec{\mathbf{x}})$. Thus, success can be expected only when the initial guess of $c(\vec{\mathbf{x}})$ is very good. What this means exactly is complicated and poorly understood at the moment.

The most powerful idea that has emerged in inversion in complex media is that of *separation of scales*. We already introduced it in Eq. (22), where we distinguished between the background speed $c(\vec{\mathbf{x}})$ (assumed smooth there) and the reflectivity ρ. We extend it in the next section to media with microscale. Reverse time migration is designed to estimate ρ, and as we explained in the previous section, it is based on the linearization $\mathcal{M}$ of the forward mapping $\mathcal{F}$. This linearization works well in many cases, but when there are strong reflecting interfaces like air-solid boundaries, the data contain strong multiply scattered echoes which cause image artifacts. See for example [36] for data processing algorithms intended to suppress such multiple reflections. In any case, migration imaging can work well only when we know the background velocity, which determines the wave propagation between the reflection events. When we have it wrong, the migration images are not well focused and the reflectivity is mapped to wrong places. The estimation of the background speed is called *velocity analysis* or *inversion*. We do not discuss it here, but we note that it is strongly connected to migration imaging [4]. On one hand, the quality of the migration images depends on knowing the background speed, and on the other hand this speed can be estimated by comparing migration images formed with various data subsets [14]. This idea is behind the iterative migration-velocity analysis inversion methods that are popular in fields like reflection seismology [4].

Cluttered Media

Imaging becomes complicated in media with rapidly fluctuating wave speed $c(\vec{\mathbf{x}})$ due to numerous small inhomogeneities. Because the array data is necessarily band limited, it is not possible to recover all these fluctuations in detail as part of the inversion. This motivates us to separate the wave speed in two parts: the "macrostructure" which can in principle be recovered in inversion, and the "microstructure" or "clutter" which is the uncertain part of $c(\vec{\mathbf{x}})$. What we call macrostructure is the smooth part $\bar{c}(\vec{\mathbf{x}})$ of the wave speed assumed known, and a sparse set of its large discontinuities modeled by the reflectivity $\rho(\vec{\mathbf{x}})$. Strictly speaking the reflectivity is a fine scale feature of $c(\vec{\mathbf{x}})$, but it can be estimated from the data so we think of it as part of the macrostructure. The assumption that $\bar{c}$ is known is because we do not consider velocity estimation. We refer to [17] for an example of robust velocity estimation in complex media with microstructure.

The separation of scales is now

$$\frac{1}{c^2(\vec{\mathbf{x}})} = \frac{1}{\bar{c}^2(\vec{\mathbf{x}})}\left[1 + \sigma\mu\left(\frac{\vec{\mathbf{x}}}{\ell}\right)\right] + \frac{\rho(\vec{\mathbf{x}})}{c_o^2}, \tag{73}$$

with $\mu(\vec{\mathbf{x}})$ representing the uncertain microstructure or clutter, the rapid fluctuations of $c(\vec{\mathbf{x}})$ on scales ℓ comparable to the wavelength or less than that. The amplitude σ of $\mu(\vec{\mathbf{x}})$ is typically small, meaning that an inhomogeneity by itself is a weak scatterer with respect to ρ. It is the cumulative effect of the inhomogeneities that becomes significant as the waves travel deeper in cluttered media. See for example the illustration in Fig. 6, where we have the same setup as in Fig. 4, but the wave speed has rapid fluctuations, as shown on the left. The amplitude of these fluctuations is small in comparison with the reflectivity of the three point-like scatterers modeled in the simulations as "sound soft," with the acoustic pressure vanishing at their boundary. These three scatterers cause strong echoes that are somewhat visible in the noisy traces, with arrivals lying on curves that are perturbations of the hyperbole in Fig. 4. But there are many other echoes recorded before and after these arrivals, due to multiple scattering by the inhomogeneities. They are more significant than the waves scattered between the reflectors in Fig. 4, so we will make a big mistake if we neglect the clutter. Much worse than using the Born approximation (linearization) with respect to ρ.

The Random Model

The clutter (microstructure) is uncertain, so we model it mathematically as a *random process* $\mu(\vec{\mathbf{x}})$, a collection of random variables parametrized by $\vec{\mathbf{x}} \in \mathbb{R}^n$. The macrostructure modeled with $\bar{c}$ and ρ in (73) is the *deterministic part* of $c(\vec{\mathbf{x}})$ because it is feasible to recover it from the array data. The wave field $p(t, \vec{\mathbf{x}}_r, \vec{\mathbf{x}}_s)$ at the array is a random process parameterized by time, the receiver

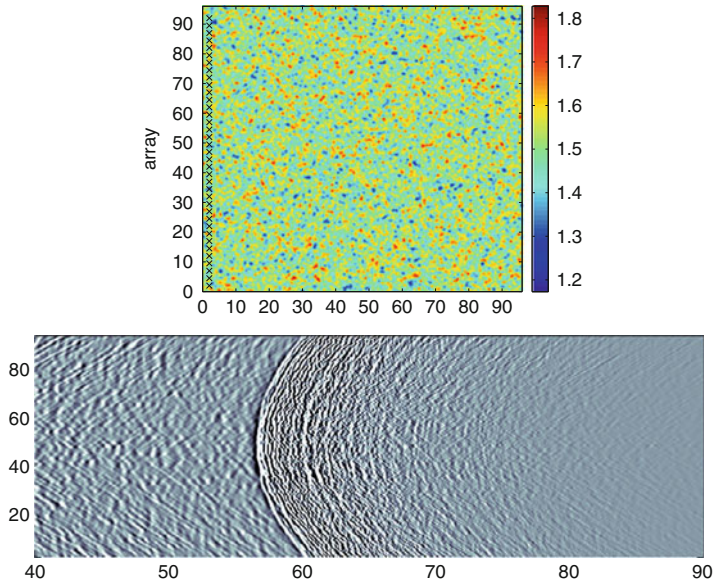


Fig. 6 Plot of the active array data traces for three reflectors in a heterogeneous medium. The abscissa is time in μs and the ordinate is receiver location in λ_o. The setup of the simulation is shown on the *left* where the units in the color bar are Km/s

and source locations. The data $\{d(t, \vec{\mathbf{x}}_r, \vec{\mathbf{x}}_s)\}$ are *one realization of this process*. If we had measurements in many media we could superpose them to approximate the statistical average (expectation) $\mathbb{E}[p(t, \vec{\mathbf{x}}_r, \vec{\mathbf{x}}_s)]$ which we call the *coherent part* of the data. This is useful in imaging because we can relate it in a robust way to the unknown reflectivity. But in reality there is only one medium, with one microstructure given by one realization of the random $\mu(\vec{\mathbf{x}})$. Thus, we do not have direct access to the coherent part of the data and must deal with its uncertain, *incoherent part* $d(t, \vec{\mathbf{x}}_r, \vec{\mathbf{x}}_s) - \mathbb{E}[p(t, \vec{\mathbf{x}}_r, \vec{\mathbf{x}}_s)]$. This impedes imaging and causes uncertainty in the reconstructions.

The reflectivity ρ is deterministic, but the wave field $p(t, \vec{\mathbf{x}}_r, \vec{\mathbf{x}}_s)$ is random so we may view the reconstructions of ρ in a probabilistic setting. A reconstruction is a mathematical model of an imaging function which involves the random field $p(t, \vec{\mathbf{x}}_r, \vec{\mathbf{x}}_s)$, so it is a random processes. We can compute only one realization of this process because we have only one realization of the random wave field, the data $\{d(t, \vec{\mathbf{x}}_r, \vec{\mathbf{x}}_s)\}$. This realization is what we call *an image*.

We study the focusing of imaging functions in random media by looking first at $\mathbb{E}[\mathcal{J}(\vec{\mathbf{y}}^s)]$ and then at its SNR at the peaks. In the case of imaging functions like reverse time migration, which are linear in $\{p(t, \vec{\mathbf{x}}_r, \vec{\mathbf{x}}_s)\}$, the expectation of the image is determined by the coherent part of the data $\mathbb{E}[p(t, \vec{\mathbf{x}}_r, \vec{\mathbf{x}}_s)]$. This is not the same as what we would measure in a medium without clutter. Multiple scattering leads to *loss of coherence of the waves*, which mani-

fests as an *exponential decay* of $\mathbb{E}[p(t, \vec{\mathbf{x}}_r, \vec{\mathbf{x}}_s)]$ with the distance of propagation. The length scale of exponential decay is called *the scattering mean-free path* and it depends on many factors like the wavelength, the range of propagation, and the statistics of $\mu(\vec{\mathbf{x}})$. Energy is conserved, so the incoherent part of the field, the random fluctuations, gain strength. We see in Fig. 6 a realization of such fluctuations, the many echoes recorded before and after the coherent arrivals.

Time Reversal Is Not Imaging

We already saw that, superficially, there is a connection between imaging and the time reversal process. This is why they are often studied side by side. But the *two are not the same at all in cluttered media*.

Time reversal works better in clutter [6, 19, 27] because scattering by the inhomogeneities spreads out the time reversed wave field captured and then re-emitted at the array, as if it came from an array of *effective aperture* $a_e > a$. The more scattering is, the larger the effective aperture, meaning that good focusing in cross-range, with resolution of the order $\lambda_o L/a_e$, can be achieved with very small arrays [19, 20]. Moreover, the refocusing is *statistically stabile* under generic assumptions [27]. This is due to the time reversability of the wave equation and the fact that the time reversed field propagates to the source in the *same medium* it came from.

In migration imaging the back-propagation is done analytically or numerically in a surrogate medium with wave speed $\bar{c}$ because we do not know the microscale of the real medium. Migration works well only when the clutter effects are weak, as shown explicitly in Sect. 7. More precisely, when the array is closer than a *scattering mean-free path* from the imaging region. Otherwise the coherent part of the data, the "signal," is exponentially damped due to scattering in clutter and the incoherent fluctuations, and the "clutter noise" is significant. The images are difficult to interpret and change unpredictably with the realization of clutter. The sole mechanism of dealing with the "clutter noise" in migration is the summation over the sensors. If we had only additive and uncorrelated noise like in section "Robustness to Additive Noise," it would average out approximately for arrays with many sensors like in the law of large numbers. But clutter noise has persistent correlations over the sensors and over frequencies, and it cannot be averaged out by summation. More involved data processing is needed to mitigate it.

A natural question arises: Could we improve the migration results if we knew the statistics of clutter? Could we just create a realization of the clutter with the given statistics and then back-propagate the data in it instead of the medium with speed $\bar{c}$? The answer is no. In fact we would do much worse than back-propagating in the smooth medium because we would essentially double the distance traveled by the waves in clutter. *The clutter effects are undone only when the waves go back through the same medium* as is the case in the time reversal process.

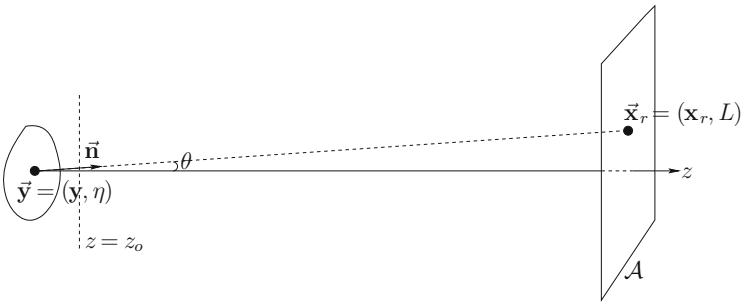


Fig. 7 System of coordinates and setup for imaging with passive arrays in random media

5 The Random Travel Time Model of Wave Propagation

In this section we describe a geometrical optics model of wave propagation in random media. It allows an explicit and self-contained analysis of resolution and statistical stability of images, and yet it captures canonical scattering effects on the wavefront. That is to say, the first two statistical moments of the wave field, which describe the loss of coherence and decorrelation of the waves due to scattering in the random medium, look the same as those derived from more sophisticated models. This makes the random travel time model a good testbed for studying imaging with partially coherent waves in random media.

Consider the setup illustrated in Fig. 7, with a passive array at range L from the source and the range axis z orthogonal to the aperture $\mathcal{A} = \left[-\frac{a}{2}, \frac{a}{2}\right] \times \left[-\frac{a}{2}, \frac{a}{2}\right]$. We use henceforth the notation $\vec{\mathbf{x}}_r = (\mathbf{x}_r, L)$ for the points in the array with cross-range components $\mathbf{x}_r \in \mathcal{A}$, and $\vec{\mathbf{y}} = (\mathbf{y}, \eta)$ for the points in the support of the source, with range η and cross-range vector $\mathbf{y}$.

We are interested in the outgoing Green's function $\hat{G}(\omega, \vec{\mathbf{x}}_r, \vec{\mathbf{y}})$ of the Helmholtz equation in an medium with random wave speed $c(\vec{\mathbf{x}})$ of the form (73) with fluctuations around the constant value $\bar{c} = c_o$. The fluctuations are modeled by the random process μ assumed stationary and twice differentiable with bounded derivatives, almost surely. It has mean zero

$$\mathbb{E}[\mu(\vec{\mathbf{u}})] = 0, \tag{74}$$

and an integrable covariance

$$\mathcal{R}(\vec{\mathbf{u}}) = \mathbb{E}\left[\mu(\vec{\mathbf{u}} + \vec{\mathbf{u}}')\mu(\vec{\mathbf{u}}')\right], \tag{75}$$

normalized by

$$\mathcal{R}(0) = 1, \qquad \int_{\mathbb{R}^3} d\vec{\mathbf{u}}\, \mathcal{R}(\vec{\mathbf{u}}) = O(1). \tag{76}$$

The length scale ℓ in (73) is the correlation length, the typical size of the inhomogeneities, and $\sigma \ll 1$ models the amplitude of the weak fluctuations of $c(\vec{\mathbf{x}})$. The fluctuations are isotropic, so the correlation length is the same in all directions, and the covariance depends on the Euclidian norm of its argument $\mathcal{R}(\vec{\mathbf{u}}) = \mathcal{R}(|\vec{\mathbf{u}}|)$. Any covariance will do, but to simplify the calculation of the statistical moments of the Green's function we take the Gaussian covariance

$$\mathcal{R}(|\vec{\mathbf{u}}|) = e^{-\frac{|\vec{\mathbf{u}}|^2}{2}}. \tag{77}$$

In the random travel time model the Green's function is approximated by

$$\hat{G}(\omega, \vec{\mathbf{x}}_r, \vec{\mathbf{y}}) \approx \frac{e^{i\omega\left[\tau_o(\vec{\mathbf{x}}_r, \vec{\mathbf{y}}) + \nu_\tau(\vec{\mathbf{x}}_r, \vec{\mathbf{y}})\right]}}{4\pi |\vec{\mathbf{x}}_r - \vec{\mathbf{y}}|}, \tag{78}$$

with reference travel time τ_o given by (63) perturbed by the random process $\nu_\tau(\vec{\mathbf{x}}_r, \vec{\mathbf{y}})$ obtained by integrating μ along the straight ray from $\vec{\mathbf{y}}$ to $\vec{\mathbf{x}}_r$,

$$\nu_\tau(\vec{\mathbf{x}}_r, \vec{\mathbf{y}}) = \frac{\sigma |\vec{\mathbf{x}}_r - \vec{\mathbf{y}}|}{2c_o} \int_0^1 dt\, \mu\left(\frac{(1-t)\vec{\mathbf{y}}}{\ell} + \frac{t\vec{\mathbf{x}}_r}{\ell}\right). \tag{79}$$

We refer to [28] for the derivation of the model, which holds in the asymptotic regime with separation of scales defined by

$$\lambda \ll \ell \lesssim L, \tag{80}$$

and

$$\sigma \ll \min\left\{\frac{\sqrt{\ell\lambda}}{L}, \left(\frac{\ell}{L}\right)^{\frac{3}{2}}\right\}. \tag{81}$$

The high frequency assumption (80) is natural for the geometrical optics approximation. Assumption (81) guarantees that ray bending is negligible from the source to the array, and the geometrical spreading factor (the amplitude of the Green's function) is approximately the same as in the homogeneous medium. It also gives that the travel time $\tau(\vec{\mathbf{x}}_r, \vec{\mathbf{y}})$, the solution of the eikonal equation in the random medium, is given by

$$\tau(\vec{\mathbf{x}}_r, \vec{\mathbf{y}}) = \tau_o(\vec{\mathbf{x}}_r, \vec{\mathbf{y}}) + \nu_\tau(\vec{\mathbf{x}}_r, \vec{\mathbf{y}}) + \delta\tau, \tag{82}$$

with negligible remainder $\delta\tau$ satisfying $|\omega\, \delta\tau| \ll 1$.

Remark 1. Referring to the traces displayed in Fig. 6, the random travel time model coupled with the Born approximation describes the echoes from the reflectors, but not the many arrivals that occur before and after them. These arrivals are incoherent

waves that do not contribute to the resolution analysis of coherent imaging methods, as long as they do not dominate the coherent arrivals. Imaging does not work when the incoherent echoes dominate unless we filter them out as explained for example in [1,2,10,13].

Long Range Scaling and Gaussian Statistics

We rewrite (79) as

$$\nu_\tau(\vec{\mathbf{x}}_r, \vec{\mathbf{y}}) = \frac{(2\pi)^{1/4}\sigma\sqrt{\mathcal{L}\ell}}{2c_o}\nu(\vec{\mathbf{x}}_r, \vec{\mathbf{y}}), \tag{83}$$

in terms of the random process

$$\nu(\vec{\mathbf{x}}_r, \vec{\mathbf{y}}) = \frac{1}{(2\pi)^{1/4}\sqrt{\ell/\mathcal{L}}}\int_0^1 dt\, \mu\left(\frac{\vec{\mathbf{y}}}{\ell} + \frac{t\vec{\mathbf{n}}}{\ell/\mathcal{L}}\right), \tag{84}$$

where $\mathcal{L} = |\vec{\mathbf{x}}_r - \vec{\mathbf{y}}|$ and $\vec{\mathbf{n}} = \frac{\vec{\mathbf{x}}_r - \vec{\mathbf{y}}}{\mathcal{L}}$ is the unit vector from the point $\vec{\mathbf{y}}$ in the source to the receiver at $\vec{\mathbf{x}}_r$. The normalization by $(2\pi)^{1/4}$ is convenient in the calculation of the statistical moments of the Green's function, for the choice (77) of the covariance of the fluctuations.

We are interested in a long range regime where the waves interact with many inhomogeneities as they travel from the source to the array

$$\ell \ll \mathcal{L} = O(L), \tag{85}$$

and the random phase in (78) is large. We know from the central limit theorem that as $\mathcal{L}/\ell \to \infty$, the right-hand side in (84) converges in distribution to a Gaussian process. Thus, the long range assumption (85) allows us to approximate $\nu(\vec{\mathbf{x}}_r, \vec{\mathbf{y}})$ by a Gaussian process, with mean zero and covariance calculated in Appendix 1,

$$\mathbb{E}\left[\nu(\vec{\mathbf{x}}, \vec{\mathbf{y}})\nu(\vec{\mathbf{x}}', \vec{\mathbf{y}}')\right] \approx \int_{-\infty}^{\infty}\frac{d\tilde{t}}{\sqrt{2\pi}}\int_0^1 dt\, \mathcal{R}\left[\tilde{t}\,\vec{\mathbf{n}} + \frac{(1-t)(\vec{\mathbf{y}}' - \vec{\mathbf{y}}) + t(\vec{\mathbf{x}}' - \vec{\mathbf{x}})}{\ell}\right]. \tag{86}$$

Moreover, we can estimate the magnitude of the random phase in (78) by

$$\omega\nu_\tau(\vec{\mathbf{x}}_r, \vec{\mathbf{y}}) = O\left(\frac{\sigma\sqrt{L\ell}}{\lambda}\right), \tag{87}$$

and ask that it be large

$$\frac{\sigma\sqrt{L\ell}}{\lambda} \gg 1, \tag{88}$$

so that the random medium has a significant effect on the wavefront, and imaging becomes difficult.

When we compare the two assumptions (81) and (88) on σ, we see that they are consistent if $\ell \gg \sqrt{\mathcal{L}\lambda}$ because then we have

$$\frac{\lambda}{\sqrt{L\ell}} \ll \min\left\{\frac{\sqrt{\lambda\ell}}{L}, \left(\frac{\ell}{L}\right)^{\frac{3}{2}}\right\} = \frac{\sqrt{\lambda\ell}}{L}.$$

Therefore, our regime is defined by the length scale separation

$$\sqrt{\lambda L} \ll \ell \ll L, \tag{89}$$

and the amplitude σ of the fluctuations satisfying

$$\frac{\lambda}{\sqrt{L\ell}} \ll \sigma \ll \frac{\sqrt{\lambda\ell}}{L}. \tag{90}$$

Statistical Moments

Since ν_τ is approximately Gaussian, we can estimate the moments of the Green's function in terms of the covariance (86). The calculations simplify for the Gaussian covariance (77), where (86) becomes

$$\mathbb{E}\left[\nu(\vec{\mathbf{x}},\vec{\mathbf{y}})\nu(\vec{\mathbf{x}}',\vec{\mathbf{y}}')\right] \approx \int_0^1 dt\, e^{-\frac{|t(\mathbf{x}'-\mathbf{x})_\perp + (1-t)(\mathbf{y}'-\mathbf{y})_\perp|^2}{2\ell^2}} \approx \int_0^1 dt\, e^{-\frac{|t(\mathbf{x}'-\mathbf{x}) + (1-t)(\mathbf{y}'-\mathbf{y})|^2}{2\ell^2}}. \tag{91}$$

This follows by direct integration over $\tilde{t}$ in (86), where $(\mathbf{x}'-\mathbf{x})_\perp$ and $(\mathbf{y}'-\mathbf{y})_\perp$ are the projection of the vectors $\vec{\mathbf{x}}-\vec{\mathbf{x}}'$ and $\vec{\mathbf{y}}'-\vec{\mathbf{y}}$ on the plane orthogonal to the ray direction $\vec{\mathbf{n}}$. The second approximation in (91) replaces these projections by the cross-range components of $\vec{\mathbf{x}}-\vec{\mathbf{x}}'$ and $\vec{\mathbf{y}}'-\vec{\mathbf{y}}$, and holds under the following scaling assumptions on the array aperture and support of the source

$$\frac{|(\vec{\mathbf{x}}'-\vec{\mathbf{x}})\cdot\vec{\mathbf{n}}|}{\ell} = \frac{|\mathbf{x}'-\mathbf{x}|}{\ell}\sin\theta = O\left(\frac{a^2}{\ell L}\right) \ll 1, \tag{92}$$

$$\frac{|\mathbf{y}'-\mathbf{y}|}{\ell}\sin\theta \approx \frac{|\mathbf{y}'-\mathbf{y}|}{\ell}\frac{a}{L} \ll 1, \qquad \frac{|\eta-\eta'|}{\ell}\cos\theta \approx \frac{|\eta'-\eta|}{\ell} \ll 1. \tag{93}$$

Loss of Coherence

Using (91) in the formula

$$\mathbb{E}\left[e^{i\omega\nu_\tau(\vec{\mathbf{x}},\vec{\mathbf{y}})}\right] \approx e^{-\frac{\omega^2}{2}\mathbb{E}\left[\nu_\tau^2(\vec{\mathbf{x}},\vec{\mathbf{y}})\right]}, \tag{94}$$

which is approximate because ν_τ is only approximately Gaussian, we obtain that

$$\mathbb{E}\left[\hat{G}(\omega,\vec{\mathbf{x}}_r,\vec{\mathbf{y}})\right] \approx \hat{G}_o(\omega,\vec{\mathbf{x}}_r,\vec{\mathbf{y}})e^{-\frac{\mathcal{L}}{\mathcal{S}(\omega)}}\,, \tag{95}$$

where $\hat{G}_o(\omega,\vec{\mathbf{x}}_r,\vec{\mathbf{y}}) = \frac{e^{ik\mathcal{L}}}{4\pi\mathcal{L}}$ is the Green's function in the homogeneous medium and

$$\mathcal{S}(\omega) = \frac{4}{\sqrt{2\pi}\,k^2\sigma^2\ell}\,, \qquad k = \frac{\omega}{c_o}. \tag{96}$$

This result shows that the coherent part of $\hat{G}$, its expectation, is smaller than $\hat{G}_o$. In fact, it is much smaller under our scaling assumption (90), which gives

$$\frac{\mathcal{L}}{\mathcal{S}(\omega)} = O\left(\frac{\sigma^2\ell L}{\lambda^2}\right) \gg 1\,. \tag{97}$$

Since

$$\mathbb{E}\left[|\hat{G}(\omega,\vec{\mathbf{x}}_r,\vec{\mathbf{y}})|^2\right] = |\hat{G}_o(\omega,\vec{\mathbf{x}}_r,\vec{\mathbf{y}})|^2 = \frac{1}{(4\pi\mathcal{L})^2}\,, \tag{98}$$

we see that the standard deviation of the field dominates its mean

$$\sqrt{\mathbb{E}\left[|\hat{G}(\omega,\vec{\mathbf{x}}_r,\vec{\mathbf{y}})|^2\right] - \left|\mathbb{E}\left[\hat{G}(\omega,\vec{\mathbf{x}}_r,\vec{\mathbf{y}})\right]\right|^2} \approx \left|\hat{G}_o(\omega,\vec{\mathbf{x}}_r,\vec{\mathbf{y}})\right| \gg \left|\mathbb{E}\left[\hat{G}(\omega,\vec{\mathbf{x}}_r,\vec{\mathbf{y}})\right]\right|\,. \tag{99}$$

That is to say, the random phase causes large random fluctuations of $\hat{G}$.

The wave emitted from $\vec{\mathbf{y}}$ and recorded at $\vec{\mathbf{x}}_r$ is given by the Fourier synthesis

$$p(t,\vec{\mathbf{x}}_r;\vec{\mathbf{y}}) = \int_{-\infty}^{\infty}\frac{d\omega}{2\pi}\hat{f}(\omega)\hat{G}(\omega,\vec{\mathbf{x}}_r,\vec{\mathbf{y}})e^{-i\omega t} \approx \frac{f\left[t-\tau_o(\vec{\mathbf{x}}_r,\vec{\mathbf{y}})-\nu_\tau(\vec{\mathbf{x}}_r,\vec{\mathbf{y}})\right]}{4\pi L}\,. \tag{100}$$

It is similar to the wave recorded in the homogeneous medium, but the arrival time of the pulse f fluctuates as described by the process ν_τ. When we take the expectation of (100) we calculate the envelope of the pulses in (100), and obtain a deformed and damped signal centered at time $\tau_o(\vec{\mathbf{x}}_r,\vec{\mathbf{y}})$. Indeed, Eqs. (95) and (100) give that

$$\mathbb{E}\left[p(t,\vec{\mathbf{x}}_r;\vec{\mathbf{y}})\right] \approx \int_{-\infty}^{\infty}\frac{d\omega}{2\pi}\frac{\hat{f}(\omega)}{4\pi\mathcal{L}}e^{-\frac{\omega^2}{2\Omega^2}-i\omega\left[t-\tau_o(\vec{\mathbf{x}}_r,\vec{\mathbf{y}})\right]} = \frac{f_\Omega\left[t-\tau_o(\vec{\mathbf{x}}_r,\vec{\mathbf{y}})\right]}{4\pi\mathcal{L}}\,, \tag{101}$$

where we defined the deformed and damped pulse

$$f_\Omega(t) = \int_{-\infty}^{\infty}\frac{d\omega}{2\pi}\hat{f}(\omega)e^{-\frac{\omega^2}{2\Omega^2}-i\omega t} = \sqrt{2\pi}\,f(t)\star_t e^{-\frac{\Omega^2 t^2}{2}}\,, \tag{102}$$

and let

$$\Omega^2 = \frac{4c_o^2}{\sqrt{2\pi}\sigma^2\ell L} \approx \frac{\omega^2 \mathcal{S}(\omega)}{\mathcal{L}} \ll \omega^2 . \tag{103}$$

For example, when the emitted pulse is the Gaussian

$$f(t) = e^{-i\omega_o t} e^{-\frac{B^2 t^2}{2}} , \qquad \hat{f}(\omega) = \frac{\sqrt{2\pi}}{B} e^{-\frac{(\omega-\omega_o)^2}{2B^2}} , \tag{104}$$

the pulse f_Ω takes the form

$$f_\Omega(t) = \frac{\Omega}{\sqrt{\Omega^2 + B^2}} \exp\left[-i\left(\frac{\Omega^2}{B^2+\Omega^2}\right)\omega_o t - \frac{B^2\Omega^2 t^2}{2(\Omega^2+B^2)} - \frac{\omega_o^2}{2(\Omega^2+B^2)}\right] . \tag{105}$$

Its central frequency is shifted to the lower value

$$\left(\frac{\Omega^2}{B^2+\Omega^2}\right)\omega_o < \omega_o ,$$

its essential support is wider by a factor of $\sqrt{1 + B^2/\Omega^2}$ than that of $f(t)$, and the amplitude of the peak is smaller

$$f_\Omega(0) = \frac{\Omega}{\sqrt{\Omega^2 + B^2}} e^{-\frac{\omega_o^2}{2(\Omega^2+B^2)}} \ll f(0) = 1 .$$

Remark 2. Formula (95) and the subsequent discussion on pulse deformation capture the loss of coherence effects derived from more sophisticated models. The loss of coherence manifests as an exponential decay of the mean wave field with the distance $\mathcal{L}$ of propagation. The length scale $\mathcal{S}(\omega)$ of decay is the scattering mean-free path and it depends on the frequency and the second-order statistics (the covariance) of the fluctuations μ. The latter appears in the expression (96) of $\mathcal{S}$ as the product $\sigma^2\ell$ in the denominator.

Statistical Decorrelation of the Waves

The second moments of the Green's function are given by

$$\mathbb{E}\left[\hat{G}(\omega,\vec{\mathbf{x}},\vec{\mathbf{y}})\overline{\hat{G}}(\omega',\vec{\mathbf{x}}',\vec{\mathbf{y}}')\right] \approx \hat{G}_o(\omega,\vec{\mathbf{x}},\vec{\mathbf{y}})\overline{\hat{G}_o}(\omega',\vec{\mathbf{x}}',\vec{\mathbf{y}}')e^{-\frac{1}{2}\mathbb{E}\left[\left(\omega\nu_\tau(\vec{\mathbf{x}},\vec{\mathbf{y}})-\omega'\nu_\tau(\vec{\mathbf{x}}',\vec{\mathbf{y}}')\right)^2\right]}, \tag{106}$$

where we used again that ν_τ is approximately Gaussian. The expectation in the exponent follows from definition (83) and the expression (91) of the covariance of ν,

$$e^{-\frac{1}{2}\mathbb{E}\left[\left(\omega\nu_\tau(\vec{\mathbf{x}},\vec{\mathbf{y}})-\omega'\nu_\tau(\vec{\mathbf{x}}',\vec{\mathbf{y}}')\right)^2\right]} \approx e^{-\frac{(\omega-\omega')^2}{2\Omega^2}-\frac{\omega\omega'}{\Omega^2}\left[1-\int_0^1 dt\, e^{-\frac{|t(\mathbf{x}-\mathbf{x}')+(1-t)(\mathbf{y}-\mathbf{y}')|^2}{2\ell^2}}\right]}. \tag{107}$$

The term in the square parenthesis is non-negative, and recalling the scaling relation (103), we note that the moments are essentially zero when $|\mathbf{x}-\mathbf{x}'|$ and/or $|\mathbf{y}-\mathbf{y}'|$ are larger or equal to ℓ. Indeed, the parenthesis in the exponent in the right-hand side of (107) is order one in this case and

$$e^{-\frac{1}{2}\mathbb{E}\left[\left(\omega\nu_\tau(\vec{\mathbf{x}},\vec{\mathbf{y}})-\omega'\nu_\tau(\vec{\mathbf{x}}',\vec{\mathbf{y}}')\right)^2\right]} = O\left(e^{-\frac{\omega_0^2}{\Omega^2}}\right) \ll 1\,.$$

When the cross-range offsets are smaller than the correlation length ℓ, we can simplify the expression (107) using the series expansion of the exponential in the square bracket, and obtain that

$$\mathbb{E}\left[\hat{G}(\omega,\vec{\mathbf{x}},\vec{\mathbf{y}})\overline{\hat{G}}(\omega',\vec{\mathbf{x}}',\vec{\mathbf{y}}')\right] \approx \hat{G}_o(\omega,\vec{\mathbf{x}},\vec{\mathbf{y}})\overline{\hat{G}_o}(\omega',\vec{\mathbf{x}}',\vec{\mathbf{y}}') \; e^{-\frac{(\omega-\omega')^2}{2\Omega^2}-\frac{|\mathbf{x}-\mathbf{x}'|^2+(\mathbf{x}-\mathbf{x}')\cdot(\mathbf{y}-\mathbf{y}')+|\mathbf{y}-\mathbf{y}'|^2}{2X^2(\bar{\omega})}}. \tag{108}$$

Here $\bar{\omega} = (\omega+\omega')/2$ is the central frequency and

$$X(\bar{\omega}) = \frac{\sqrt{3}\,\Omega}{\bar{\omega}}\ell\,. \tag{109}$$

In the derivation of (109) we used that $\Omega \ll \bar{\omega}$, as stated by the scaling relation (103). The result says that the moments decay exponentially with the cross-range offsets, on the scale $X(\bar{\omega}) \ll \ell$, which is consistent with the assumption $|\mathbf{x}-\mathbf{x}'|, |\mathbf{y}-\mathbf{y}'| \ll \ell$ used in the derivation. For larger cross-range offsets, the moments are exponentially small, as noted above. This is precisely what (109) gives, so we can use the result for all cross-range offsets, with negligible errors in the analysis of imaging in the following sections.

It is clear that the maximum of (109) occurs at the same frequency $\omega' = \omega$ and the same points in the cross-range plane $\mathbf{x}' = \mathbf{x}$ and $\mathbf{y}' = \mathbf{y}$. The decay of the moments with the frequency and cross-range offsets describes the statistical decorrelation of the waves. Indeed, the correlation coefficient is defined by

$$\mathrm{Corr}\left[\hat{G}(\omega,\vec{\mathbf{x}},\vec{\mathbf{y}}),\hat{G}(\omega',\vec{\mathbf{x}}',\vec{\mathbf{y}}')\right] = \frac{\mathbb{E}\left[\hat{G}(\omega,\vec{\mathbf{x}},\vec{\mathbf{y}})\overline{\hat{G}}(\omega',\vec{\mathbf{x}}',\vec{\mathbf{y}}')\right]-\mathbb{E}\left[\hat{G}(\omega,\vec{\mathbf{x}},\vec{\mathbf{y}})\right]\mathbb{E}\left[\overline{\hat{G}}(\omega',\vec{\mathbf{x}}',\vec{\mathbf{y}}')\right]}{\sqrt{\mathrm{Var}\left[\hat{G}(\omega,\vec{\mathbf{x}},\vec{\mathbf{y}})\right]\mathrm{Var}\left[\hat{G}(\omega',\vec{\mathbf{x}}',\vec{\mathbf{y}}')\right]}},$$

with variance estimated in (99). Recalling from (95) that the mean field is exponentially small, we estimate the correlation as

$$\left|\mathrm{Corr}\left[\hat{G}(\omega,\vec{\mathbf{x}},\vec{\mathbf{y}}),\hat{G}(\omega',\vec{\mathbf{x}}',\vec{\mathbf{y}}')\right]\right| \approx e^{-\frac{(\omega-\omega')^2}{2\Omega^2}-\frac{|\mathbf{x}-\mathbf{x}'|^2+(\mathbf{x}-\mathbf{x}')\cdot(\mathbf{y}-\mathbf{y}')+|\mathbf{y}-\mathbf{y}'|^2}{2X^2(\bar{\omega})}}. \tag{110}$$

It decays with the frequency offset on the scale Ω, called the *decoherence frequency*, and with the cross-range offsets on the length scale X, called the *decorrelation length*. This scale is proportional to the wavelength, as stated in (109).

Remark 3. The expression of the moments in (109) is the same as that derived from more sophisticated models of cumulative scattering effects in random media. The detailed expressions of the decoherence frequency and decorrelation length change with the model, but Eq. (109) captures the canonical form of the second moments, and thus of the statistical decorrelation of the waves.

6 Setup for Imaging

We consider the setup for imaging with passive arrays illustrated in Fig. 7, with the range axis originating from the center of the source and the array of aperture $\mathcal{A}$ orthogonal to it. For active arrays, the excitation comes from a source in $\mathcal{A}$ and ρ is the reflectivity that we wish to estimate. In either case the array is assumed planar and square, with side length a satisfying the scaling relation (92). We let a be larger or similar to the correlation length, so that the Fresnel number Φ_a satisfies

$$\Phi_a = \frac{a^2}{\lambda L} \gtrsim \frac{\ell^2}{\lambda L} \gg 1\,. \tag{111}$$

The waves are approximately planar in regimes with smaller Fresnel number and we cannot estimate the cross-range location of the support of ρ.

We restrict the analysis to a paraxial regime, where we can write

$$\omega\tau_o(\vec{\mathbf{x}}_r,\vec{\mathbf{y}}) = k\sqrt{(L-\eta)^2+|\mathbf{x}_r-\mathbf{y}|^2} \approx k\left[L-\eta+\frac{|\mathbf{x}_r|^2}{2L}-\frac{\mathbf{x}_r\cdot\mathbf{y}}{L}\right]. \tag{112}$$

This holds for apertures satisfying $\frac{a^4}{\lambda L^3} \ll 1$, and supports of ρ of radius R in cross-range and R_η in range, satisfying $\frac{R^2}{\lambda L} \ll 1$ and $R_\eta \ll \frac{L^2\lambda}{a^2}$. Merging all our assumptions on a, we obtain that

$$\sqrt{\lambda L} \ll a \ll \min\left\{\sqrt{\ell L},\left(\lambda L^3\right)^{1/4}\right\} = \left(\lambda L^3\right)^{1/4}, \tag{113}$$

where the last equality holds in our regime defined by (89) because

$$\frac{\sqrt{\ell L}}{(\lambda L^3)^{1/4}} = \left(\frac{\ell}{\sqrt{\lambda L}}\right)^{1/2} \gg 1\,.$$

The summary of the assumptions on the cross-range support of ρ is

$$R \ll \min\left\{\sqrt{\lambda L}, \frac{\ell L}{a}\right\} = \sqrt{\lambda L}\,, \tag{114}$$

where we recalled (93) and used that

$$\frac{\ell L}{a\sqrt{\lambda L}} = \frac{\ell}{\sqrt{\lambda L}}\frac{L}{a} \gg 1\,.$$

The range support of ρ satisfies

$$R_\eta \ll \min\left\{\ell, \frac{L^2\lambda}{a^2}\right\}\,. \tag{115}$$

We take the complex-valued Gaussian pulse (104) for the sake of explicit calculations. Its bandwidth B satisfies

$$\frac{c_o}{L} \ll B \ll \omega_o\,. \tag{116}$$

The first inequality ensures that $f(t)$ is a pulse of small temporal support, so we can determine the travel times $\tau_o(\vec{\mathbf{x}}_r, \vec{\mathbf{y}}) = O(L/c_o)$ from the measurements of the acoustic pressure $p(t, \vec{\mathbf{x}}_r)$ at the array. The second inequality means that all the frequencies in the bandwidth are of the order ω_o and therefore the wavelengths satisfy $\lambda \approx \lambda_o$. The decoherence frequency and decorrelation length defined in Sect. 5 obey the scaling relations

$$\Omega \ll \omega_o\,, \qquad X(\omega) \approx X(\omega_o) \ll \ell \lesssim a\,. \tag{117}$$

To simplify the analysis of the CINT imaging function in Sect. 8, we assume in addition that the decorrelation length satisfies

$$X(\omega_o) \ll \sqrt{\lambda_o L}. \tag{118}$$

Using the definition (109) of $X(\omega_o)$, we see that this is equivalent to asking that

$$\Omega \ll \omega_o \frac{\sqrt{\lambda_o L}}{\ell} \ll \omega_o, \tag{119}$$

where the second inequality is implied by (89).

We analyze in Sects. 7 and 8 the imaging function $\mathcal{J}(\vec{\mathbf{y}}^s)$ of two methods: Kirchhoff migration and coherent interferometry. We use the same symbol for the imaging functions to simplify the notation, and define $\mathcal{J}(\vec{\mathbf{y}}^s)$ at points $\vec{\mathbf{y}}^s$ in a search domain that contains the support of the source. The size of this domain is chosen, so that we can use the paraxial approximation (112) for the travel time $\tau_o(\vec{\mathbf{x}}, \vec{\mathbf{y}}^s)$. We use for convenience the continuum aperture approximation, which is based on the assumption that the sensors in the array are close together. The approximation amounts to replacing sums over the sensors by scaled integrals over the aperture $\mathcal{A}$. The scaling factor is $N/|\mathcal{A}|$, with N the number of sensors and $|\mathcal{A}| = a^2$ the area of the aperture.

The resolution analysis of the imaging function $\mathcal{J}(\vec{\mathbf{y}}^s)$ involves two steps: In the first step we study the focusing of its expectation. Since the imaging process occurs in one medium, corresponding to one realization of the random process μ, we do not have access to $\mathbb{E}[\mathcal{J}(\vec{\mathbf{y}}^s)]$. But if the imaging were robust, the result would look almost the same independent of the realization, so $\mathcal{J}(\vec{\mathbf{y}}^s)$ would be close to its expectation, which is why we study it. The second step of the resolution analysis is the assessment of the statistical stability of $\mathcal{J}(\vec{\mathbf{y}}^s)$. It amounts to calculating the SNR of the image at its peaks.

Most of our discussion is about imaging with passive arrays. The extension to active arrays is straightforward for our random travel time model combined with the Born approximation.

7 Migration Imaging

In this section we analyze migration imaging with passive arrays. We consider only the Kirchhoff migration imaging function because it is essentially the same as reverse time migration in the paraxial regime described in Sect. 6. It is given by

$$\mathcal{J}(\vec{\mathbf{y}}^s) = \sum_{r=1}^{N} p\left(\tau_o(\vec{\mathbf{x}}_r, \vec{\mathbf{y}}^s), \vec{\mathbf{x}}_r\right) \approx \frac{N}{|\mathcal{A}|} \int_{\mathcal{A}} d\mathbf{x}\, p\left(\tau_o(\vec{\mathbf{x}}, \vec{\mathbf{y}}^s), \vec{\mathbf{x}}\right), \qquad \vec{\mathbf{x}} = (\mathbf{x}, L), \tag{120}$$

where $p(t, \vec{\mathbf{x}})$ is the pressure field modeled by

$$p(t, \vec{\mathbf{x}}) = \int_{\mathbb{R}^3} d\vec{\mathbf{y}} \rho(\vec{\mathbf{y}}) p(t, \vec{\mathbf{x}}; \vec{\mathbf{y}}), \tag{121}$$

in terms of the spatial source density $\rho(\vec{\mathbf{y}})$ and the field $p(t, \vec{\mathbf{x}}; \vec{\mathbf{y}})$ due to a point source at $\vec{\mathbf{y}}$. The latter is defined in (100), and we rewrite it here using the simplification of the Green's function

$$\hat{G}(\omega, \vec{\mathbf{x}}_r, \vec{\mathbf{y}}) \approx \frac{e^{i\omega[\tau_o(\vec{\mathbf{x}}_r, \vec{\mathbf{y}}) + \nu_\tau(\vec{\mathbf{x}}_r, \vec{\mathbf{y}})]}}{4\pi L} \approx \frac{1}{4\pi L} e^{ik\left(L - \eta + \frac{|\mathbf{x}_r|^2}{2L} - \frac{\mathbf{x}_r \cdot \mathbf{y}}{L}\right) + i\omega\nu_\tau(\vec{\mathbf{x}}_r, \vec{\mathbf{y}})}, \tag{122}$$

that holds under the scaling assumptions (113) and (114). The wave field due to the point source is given by

$$p(t,\vec{\mathbf{x}};\vec{\mathbf{y}}) \approx \int_{-\infty}^{\infty} d\omega\, \frac{\hat{f}(\omega)}{8\pi^2 L}\, e^{i\omega[\tau_o(\vec{\mathbf{x}},\vec{\mathbf{y}})-t]+i\omega v_\tau(\vec{\mathbf{x}},\vec{\mathbf{y}})} \approx$$
$$\int_{-\infty}^{\infty} d\omega\, \frac{\hat{f}(\omega)}{8\pi^2 L}\, e^{ik\left(L-\eta+\frac{|\mathbf{x}_r|^2}{2L}-\frac{\mathbf{x}_r\cdot\mathbf{y}}{L}\right)+i\omega v_\tau(\vec{\mathbf{x}},\vec{\mathbf{y}})-i\omega t}\,. \tag{123}$$

The Expectation

Taking the expectation in (120) and using (121), we obtain that

$$\mathbb{E}\left[\mathcal{J}(\vec{\mathbf{y}}^s)\right] = \int_{\mathbb{R}^3} d\vec{\mathbf{y}}\rho(\vec{\mathbf{y}})\mathbb{E}\left[\mathcal{J}(\vec{\mathbf{y}}^s;\vec{\mathbf{y}})\right]\,, \tag{124}$$

with mean point spread function

$$\mathbb{E}\left[\mathcal{J}(\vec{\mathbf{y}}^s;\vec{\mathbf{y}})\right] \approx \frac{N}{|\mathcal{A}|}\int_{\mathcal{A}} d\mathbf{x}\,\mathbb{E}\left[p\left(\tau_o(\vec{\mathbf{x}},\vec{\mathbf{y}}^s),\vec{\mathbf{x}};\vec{\mathbf{y}}\right)\right]\,. \tag{125}$$

The expectation of p is given by (101), and using the approximation (112) of the travel time we write

$$\tau_o(\vec{\mathbf{x}},\vec{\mathbf{y}}^s) - \tau_o(\vec{\mathbf{x}},\vec{\mathbf{y}}) \approx \frac{1}{c_o}\left[\eta^s - \eta + \frac{(\mathbf{y}^s-\mathbf{y})\cdot\mathbf{x}}{L}\right], \tag{126}$$

for $\vec{\mathbf{y}}^s = (\mathbf{y}^s,\eta^s)$. Equation (125) becomes

$$\mathbb{E}\left[\mathcal{J}(\vec{\mathbf{y}}^s;\vec{\mathbf{y}})\right] \approx \frac{N\, e^{-\frac{\omega_o^2}{2(B^2+\Omega^2)}}}{4\pi L\sqrt{B^2/\Omega^2+1}}$$
$$\int_{\mathcal{A}}\frac{d\mathbf{x}}{|\mathcal{A}|}\, e^{-\frac{i\left[k_o(\eta^s-\eta)+\frac{k_o(\mathbf{y}^s-\mathbf{y})\cdot\mathbf{x}}{L}\right]}{(B^2/\Omega^2+1)}-\frac{B^2\left[(\eta^s-\eta)+\frac{(\mathbf{y}^s-\mathbf{y})\cdot\mathbf{x}}{L}\right]^2}{2c_o^2(B^2/\Omega^2+1)}}\,, \tag{127}$$

where $k_o = \omega_o/c_o$ is the central wave number, and it is not difficult to see that the mean point spread function peaks at $\vec{\mathbf{y}} = (\mathbf{y},\eta)$, as it should.

To determine the range resolution, we evaluate (127) at $\mathbf{y}^s = \mathbf{y}$

$$\mathbb{E}\left[\mathcal{J}(\vec{\mathbf{y}}^s = (\mathbf{y},\eta^s);\vec{\mathbf{y}})\right] \approx \frac{N\, e^{-\frac{\omega_o^2}{2(B^2+\Omega^2)}}}{4\pi L\sqrt{B^2/\Omega^2+1}}\, e^{-\frac{ik_o(\eta^s-\eta)}{(B^2/\Omega^2+1)}-\frac{B^2(\eta^s-\eta)^2}{2c_o^2(B^2/\Omega^2+1)}}\,, \tag{128}$$

and estimate the range resolution in terms of the standard deviation of the Gaussian

$$|\eta^s - \eta| \lesssim \frac{c_o}{B}\sqrt{B^2/\Omega^2 + 1}\,. \tag{129}$$

The result in the homogeneous medium corresponds to letting $\Omega \to \infty$ in this equation, in which case the range resolution would be c_o/B, as stated in section "Imaging in Smooth and Known Media." In our regime the resolution is worse due to the broadening of the pulse f_Ω noted in section "Statistical Moments."

The cross-range resolution is estimated from

$$\begin{aligned}
\mathbb{E}\left[\mathcal{J}(\vec{\mathbf{y}}^s = (\mathbf{y}^s, \eta); \vec{\mathbf{y}})\right] &\approx \frac{N\, e^{-\frac{\omega_o^2}{2(B^2/\Omega^2+1)}}}{4\pi L\sqrt{B^2+\Omega^2}} \int_{\mathcal{A}} \frac{d\mathbf{x}}{|\mathcal{A}|}\, e^{-\frac{ik_o(\mathbf{y}^s-\mathbf{y})\cdot\mathbf{x}/L}{(B^2/\Omega^2+1)} - \frac{B^2[k_o(\mathbf{y}^s-\mathbf{y})\cdot\mathbf{x}/L]^2}{2\omega_o^2(B^2/\Omega^2+1)}} \\
&\approx \frac{N\, e^{-\frac{\omega_o^2}{2(B^2+\Omega^2)}}}{4\pi L\sqrt{B^2/\Omega^2+1}} \int_{\mathcal{A}} \frac{d\mathbf{x}}{|\mathcal{A}|}\, e^{-\frac{ik_o(\mathbf{y}^s-\mathbf{y})\cdot\mathbf{x}/L}{(B^2/\Omega^2+1)}} \\
&= \frac{N\, e^{-\frac{\omega_o^2}{2(B^2+\Omega^2)}}}{4\pi L\sqrt{B^2/\Omega^2+1}} \prod_{j=1}^{2} \operatorname{sinc}\left[\frac{k_o a(\mathbf{y}^s-\mathbf{y})\cdot\mathbf{e}_j}{2L(B^2/\Omega^2+1)}\right].
\end{aligned} \tag{130}$$

The first approximation holds because B and Ω are much smaller than ω_o, and the third line follows by integration over the square aperture $\mathcal{A}$. We estimate the cross-range resolution by the support of the main peak of the sinc function

$$|(\mathbf{y}^s - \mathbf{y})\cdot\mathbf{e}_j| \lesssim \frac{\lambda_o L}{a}\left(\frac{B^2}{\Omega^2} + 1\right). \tag{131}$$

It is worse than that in the homogeneous medium of $\lambda_o L/a$ because of the shift of the central frequency of the deformed pulse $f_\Omega(t)$ to a lower value, as noted in section "Statistical Moments."

The value of the mean point spread function at the peak is given by

$$\mathbb{E}\left[\mathcal{J}(\vec{\mathbf{y}};\vec{\mathbf{y}})\right] \approx \frac{N\, e^{-\frac{\omega_o^2}{2(B^2+\Omega^2)}}}{4\pi L\sqrt{B^2/\Omega^2+1}} \ll \mathcal{J}_o(\vec{\mathbf{y}};\vec{\mathbf{y}}) = \frac{N}{4\pi L}\,. \tag{132}$$

It is much smaller than $\mathcal{J}_o(\vec{\mathbf{y}};\vec{\mathbf{y}})$, the peak of the point spread function in the homogeneous medium. Finally, the mean imaging function (124) equals the unknown density ρ integrated against the blurring kernel (127).

The SNR

To simplify the calculations, we consider only the SNR of the point spread function $\mathcal{J}(\vec{\mathbf{y}}^s;\vec{\mathbf{y}})$ evaluated at the peak $\vec{\mathbf{y}}^s = \vec{\mathbf{y}}$. We have

$$\mathcal{J}(\vec{\mathbf{y}};\vec{\mathbf{y}}) \approx N \int_{\mathcal{A}} \frac{d\mathbf{x}}{|\mathcal{A}|}\, p\left(\tau_o(\vec{\mathbf{x}},\vec{\mathbf{y}}),\vec{\mathbf{x}};\vec{\mathbf{y}}\right) \approx \frac{N\sqrt{2\pi}}{8\pi^2 L\, B|\mathcal{A}|}\int_{\mathcal{A}} d\mathbf{x} \int_{-\infty}^{\infty} d\omega\, e^{-\frac{(\omega-\omega_0)^2}{2B^2}+i\omega v_t(\vec{\mathbf{x}},\vec{\mathbf{y}})}, \tag{133}$$

where we used Eqs. (100) and (122) and the Gaussian pulse (104). The mean square follows from (109)

$$\mathbb{E}\left[\left|\mathcal{J}(\vec{\mathbf{y}};\vec{\mathbf{y}})\right|^2\right] \approx \frac{N^2}{32\pi^3 L^2 B^2|\mathcal{A}|^2}\iint_{\mathcal{A}} d\mathbf{x}d\mathbf{x}' \iint_{-\infty}^{\infty} d\omega d\omega' e^{-\frac{(\omega-\omega_0)^2}{2B^2}-\frac{(\omega'-\omega_0)^2}{2B^2}-\frac{(\omega-\omega')}{2\Omega^2}-\frac{|\mathbf{x}'-\mathbf{x}|^2}{2X^2(\omega_0)}},$$

and changing variables to

$$\bar{\omega} = \frac{\omega+\omega'}{2}, \qquad \tilde{\omega} = \omega-\omega' \quad \text{and} \quad \bar{\mathbf{x}} = \frac{\mathbf{x}+\mathbf{x}'}{2}, \qquad \tilde{\mathbf{x}} = \mathbf{x}-\mathbf{x}',$$

we obtain

$$\mathbb{E}\left[\left|\mathcal{J}(\vec{\mathbf{y}};\vec{\mathbf{y}})\right|^2\right] \approx \frac{N^2}{32\pi^3 L^2 B|\mathcal{A}|}\int_{R^2} d\tilde{\mathbf{x}}\, e^{-\frac{|\tilde{\mathbf{x}}|^2}{2X^2(\omega_0)}}\int_{-\infty}^{\infty}\frac{d\bar{\omega}}{B}e^{-\frac{(\bar{\omega}-\omega_0)^2}{B^2}} \int_{-\infty}^{\infty} d\tilde{\omega}\, e^{-\frac{\tilde{\omega}^2}{2}\left(\frac{1}{\Omega^2}+\frac{1}{2B^2}\right)} = \frac{N^2X^2}{8\pi L^2|\mathcal{A}|\sqrt{2B^2/\Omega^2+1}}. \tag{134}$$

In the first line of this equation we integrated over $\bar{\mathbf{x}}$ and used the scaling assumption (117) on the decorrelation length X to extend the integral over $\tilde{\mathbf{x}}$ to the whole plane. The second line follows by integrating the Gaussians in $\tilde{\mathbf{x}}$, $\bar{\omega}$ and $\tilde{\omega}$.

The variance of the point spread function at the peak is derived from (132) and (134)

$$\mathrm{Var}\left[\mathcal{J}(\vec{\mathbf{y}};\vec{\mathbf{y}})\right] = \mathbb{E}\left[\left|\mathcal{J}(\vec{\mathbf{y}};\vec{\mathbf{y}})\right|^2\right] - \left|\mathbb{E}\left[\mathcal{J}(\vec{\mathbf{y}};\vec{\mathbf{y}})\right]\right|^2 \approx \frac{N^2X^2}{8\pi L^2|\mathcal{A}|\sqrt{2B^2/\Omega^2+1}}, \tag{135}$$

and the SNR is given by

$$\mathrm{SNR}\left[\mathcal{J}(\vec{\mathbf{y}};\vec{\mathbf{y}})\right] = \frac{\left|\mathbb{E}\left[\mathcal{J}(\vec{\mathbf{y}};\vec{\mathbf{y}})\right]\right|}{\sqrt{\mathrm{Var}\left[\mathcal{J}(\vec{\mathbf{y}};\vec{\mathbf{y}})\right]}} \approx \frac{1}{\sqrt{2\pi}}\frac{a}{X}\frac{(2B^2/\Omega^2+1)^{1/4}}{(B^2/\Omega^2+1)^{1/2}} e^{-\frac{\omega_o^2}{2(B^2+\Omega^2)}}, \tag{136}$$

where we used that $|\mathcal{A}| = a^2$. Thus, we see that the SNR improves when we increase the aperture and the bandwidth. We restricted the bandwidth by $B \ll \omega_o$, so the exponentially small factor plays an important role in the SNR. The definition (109) of the decorrelation length gives that

$$\frac{a}{X} = O\left(\frac{a\,\omega_o}{\Omega\,\ell}\right)$$

and to achieve a large SNR we need

$$\frac{a}{\ell} \gg \frac{\Omega}{\omega_o}\exp\left[\frac{\omega_o^2}{2(B^2+\Omega^2)}\right].$$

The first factor is small, but the exponential is huge and such large apertures cannot be realized in practice. Therefore, the SNR of the Kirchhoff migration function is small and that the imaging function is not robust.

8 CINT Imaging

The statistical instability of migration imaging is due to the large random phase $\omega\nu_\tau(\vec{\mathbf{x}}_r,\vec{\mathbf{y}})$ in the Green's function, so a natural idea for improving the imaging is to cancel the random phase by data processing. We could cancel the phase exactly by working with the intensities $|\hat{p}(\omega,\vec{\mathbf{x}}_r)|^2$ of the measurements, but this is not a good idea. To see why consider a point source at $\vec{\mathbf{y}}$, so that

$$\hat{p}(\omega,\vec{\mathbf{x}}_r) = \hat{p}(\omega,\vec{\mathbf{x}}_r;\vec{\mathbf{y}}) = \hat{f}(\omega)\hat{G}(\omega,\vec{\mathbf{x}}_r,\vec{\mathbf{y}}). \tag{137}$$

The intensity is

$$|\hat{p}(\omega,\vec{\mathbf{x}}_r;\vec{\mathbf{y}})|^2 = |\hat{f}(\omega)|^2|\hat{G}(\omega,\vec{\mathbf{x}}_r,\vec{\mathbf{y}})|^2 \approx \frac{|\hat{f}(\omega)|^2}{(4\pi|\vec{\mathbf{x}}_r-\vec{\mathbf{y}}|)^2} \approx \frac{|\hat{f}(\omega)|^2}{(4\pi L)^2}$$

and it is nearly impossible to estimate the location $\vec{\mathbf{y}}$ from it because it is approximately constant across the aperture. We need to keep the deterministic phase, the travel time $\tau_o(\vec{\mathbf{x}}_r,\vec{\mathbf{y}})$, in order to estimate $\vec{\mathbf{y}}$. At the same, we should reduce the random phase $\omega\nu_\tau(\vec{\mathbf{x}}_r,\vec{\mathbf{y}})$.

The CINT imaging approach accomplishes this by migrating to $\vec{\mathbf{y}}^s$ local cross-correlations of the data instead of the data themselves. The local cross-correlations

are defined by

$$\mathcal{C}(t,\tilde{t};\vec{\mathbf{x}},\vec{\mathbf{x}}') = \int_{-\infty}^{\infty} ds\, \Omega\, \Phi\left[\Omega(t-s)\right]\; p\left(s+\tilde{t}/2,\vec{\mathbf{x}}\right)\overline{p}\left(s-\tilde{t}/2,\vec{\mathbf{x}}'\right)$$
$$= \int_{-\infty}^{\infty}\frac{d\omega}{2\pi}\int_{-\infty}^{\infty}\frac{d\tilde{\omega}}{2\pi}\,\hat{\Phi}\left(\frac{\tilde{\omega}}{\Omega}\right)\hat{p}(\omega+\tilde{\omega}/2,\vec{\mathbf{x}})\,\overline{\hat{p}}\,(\omega-\tilde{\omega}/2,\vec{\mathbf{x}}')$$
$$e^{-i\tilde{\omega}t-i\omega\tilde{t}}. \tag{138}$$

They are calculated around the time t, in a time window Φ of duration $1/\Omega$ determined by the decoherence frequency Ω with a purpose! To explain the advantage of working with the cross-correlations (138), we analyze them in section "Analysis of the Cross-Correlations for a Point Source" for the case of a imaging a point source. Then we study in sections "Resolution Analysis of the CINT Imaging Function" and "CINT Images for Passive Arrays as the Smoothed Wigner Transform" the resolution of the CINT imaging method with passive arrays. It forms an image by migrating to search points $\vec{\mathbf{y}}^s$ in the imaging domain the cross-correlations (138). Explicitly, the imaging function is given by

$$\mathcal{J}(\vec{\mathbf{y}}^s) = \sum_{r,r'=1}^{N}\int_{-\infty}^{\infty}\frac{d\omega}{2\pi}\int_{-\infty}^{\infty}\frac{d\tilde{\omega}}{2\pi}\,\Psi\left[\frac{\mathbf{x}_r-\mathbf{x}_{r'}}{X(\omega)}\right]\hat{\Phi}\left(\frac{\tilde{\omega}}{\Omega}\right)\hat{p}(\omega+\tilde{\omega}/2,\vec{\mathbf{x}}_r)$$
$$\overline{\hat{p}}\,(\omega-\tilde{\omega}/2,\vec{\mathbf{x}}_{r'})$$
$$\times\; e^{-i\tilde{\omega}\left[\frac{\tau_o(\vec{\mathbf{x}}_r,\vec{\mathbf{y}}^s)+\tau_o(\vec{\mathbf{x}}_{r'},\vec{\mathbf{y}}^s)}{2}\right]-i\omega[\tau_o(\vec{\mathbf{x}}_r,\vec{\mathbf{y}}^s)-\tau_o(\vec{\mathbf{x}}_{r'},\vec{\mathbf{y}}^s)]}, \tag{139}$$

where Ψ is another window function that keeps the cross-range offsets $\mathbf{x}_r-\mathbf{x}_{r'}$ within the distance $X(\omega)$. As was the case with the time window above, the support $X(\omega)$ is chosen equal to the decorrelation length with a purpose. Recalling from (117) that in our scaling regime $X(\omega)\approx X(\omega_o)$ and using (138), we can approximate the CINT imaging function as

$$\mathcal{J}(\vec{\mathbf{y}}^s)\approx\sum_{r,r'=1}^{N}\Psi\left[\frac{\mathbf{x}_r-\mathbf{x}_{r'}}{X(\omega_o)}\right]\mathcal{C}\left(\overline{\tau}_o(\vec{\mathbf{x}}_r,\vec{\mathbf{x}}_{r'},\vec{\mathbf{y}}^s),\tilde{\tau}_o(\vec{\mathbf{x}}_r,\vec{\mathbf{x}}_{r'},\vec{\mathbf{y}}^s);\vec{\mathbf{x}}_r,\vec{\mathbf{x}}_{r'}\right). \tag{140}$$

This is just the superposition of the migrated local cross-correlations (138) at nearby receivers. The migration amounts to evaluating the cross-correlations at the average and differences of the travel times from the receivers to the search point

$$t=\overline{\tau}_o(\vec{\mathbf{x}}_r,\vec{\mathbf{x}}_{r'},\vec{\mathbf{y}}^s)=\frac{\tau_o(\vec{\mathbf{x}}_r,\vec{\mathbf{y}}^s)+\tau_o(\vec{\mathbf{x}}_{r'},\vec{\mathbf{y}}^s)}{2},$$
$$\tilde{t}=\tilde{\tau}_o(\vec{\mathbf{x}}_r,\vec{\mathbf{x}}_{r'},\vec{\mathbf{y}}^s)=\tau_o(\vec{\mathbf{x}}_r,\vec{\mathbf{y}}^s)-\tau_o(\vec{\mathbf{x}}_{r'},\vec{\mathbf{y}}^s). \tag{141}$$

The resolution analysis below applies to any window functions, but for the sake of explicit and simpler formulas, we consider the Gaussian windows

$$\Phi(s) = e^{-s^2/2}, \qquad \Psi(\mathbf{u}) = e^{-|\mathbf{u}|^2/2}. \tag{142}$$

Analysis of the Cross-Correlations for a Point Source

The model of the Fourier transform of the measurements is given by (137), with $\hat{G}(\omega, \vec{\mathbf{x}}_r, \vec{\mathbf{y}})$ approximated by (122). Using a notation similar to (141) for the average and differences of the travel times, and the Gaussian pulse (104), we get

$$\mathcal{C}(t, \tilde{t}; \vec{\mathbf{x}}, \vec{\mathbf{x}}') \approx \frac{1}{2\pi(4\pi L)^2 B^2} \int_{-\infty}^{\infty} d\omega\, e^{-\frac{(\omega-\omega_o)^2}{B^2}} \int_{-\infty}^{\infty} d\tilde{\omega}\, \hat{\Phi}\left(\frac{\tilde{\omega}}{\Omega}\right) e^{-\frac{\tilde{\omega}^2}{4B^2} + i\omega\left[\tilde{\tau}_o(\vec{\mathbf{x}}, \vec{\mathbf{x}}', \vec{\mathbf{y}}) - \tilde{t}\right]}$$
$$\times\, e^{i\tilde{\omega}\left[\overline{\tau}_o(\vec{\mathbf{x}}, \vec{\mathbf{x}}', \vec{\mathbf{y}}) - t\right] + i(\omega+\tilde{\omega}/2)\nu_\tau(\vec{\mathbf{x}}, \vec{\mathbf{y}}) - i(\omega-\tilde{\omega}/2)\nu_\tau(\vec{\mathbf{x}}', \vec{\mathbf{y}})}. \tag{143}$$

We calculate first the coherent part of $\mathcal{C}$, its statistical expectation, to show that it peaks at times

$$t = \overline{\tau}_o(\vec{\mathbf{x}}_r, \vec{\mathbf{x}}_{r'}, \vec{\mathbf{y}}) \quad \text{and} \quad \tilde{t} = \tilde{\tau}_o(\vec{\mathbf{x}}_r, \vec{\mathbf{x}}_{r'}, \vec{\mathbf{y}}).$$

Then we estimate the SNR at the peak to assess its statistical stability with respect to the realizations of the random medium.

The expectation of (143) follows from (109)

$$\mathbb{E}\left[\mathcal{C}(t, \tilde{t}; \vec{\mathbf{x}}, \vec{\mathbf{x}}')\right] \approx \frac{1}{2\pi(4\pi L)^2 B^2}\, e^{-\frac{|\mathbf{x}-\mathbf{x}'|^2}{2X^2(\omega_o)}} \int_{-\infty}^{\infty} d\omega\, e^{-\frac{(\omega-\omega_o)^2}{B^2} + i\omega\left[\tilde{\tau}_o(\vec{\mathbf{x}}, \vec{\mathbf{x}}', \vec{\mathbf{y}}) - \tilde{t}\right]}$$
$$\int_{-\infty}^{\infty} d\tilde{\omega}\, \hat{\Phi}\left(\frac{\tilde{\omega}}{\Omega}\right) e^{-\frac{\tilde{\omega}^2}{2}\left(\frac{1}{\Omega^2} + \frac{1}{2B^2}\right) + i\tilde{\omega}\left[\overline{\tau}_o(\vec{\mathbf{x}}, \vec{\mathbf{x}}', \vec{\mathbf{y}}) - t\right]}, \tag{144}$$

where we approximated the decorrelation length by that at the central frequency ω_o, as stated in the scaling relation (117). Note that the essential support of $\tilde{\omega}$ in the second integral is limited to $|\tilde{\omega}| \lesssim \min\{\Omega, B\}$, because the pulse has finite bandwidth, and the waves decorrelate over frequency offsets that are larger than Ω. Thus, if we had a window Φ with Fourier transform supported over a larger frequency interval than Ω, it would make no difference in (144). However, we will see later that it will result in lower SNR, so we would lose in statistical stability. Could we take a smaller frequency support of $\hat{\Phi}$ than Ω? Such a choice would reduce the variance of the fluctuations of $\mathcal{C}$, but it would broaden the peak along the t axis. This follows from the evaluation of the $\tilde{\omega}$ integral in (144). The optimal choice of the frequency support is the decoherence frequency Ω, as we take it in (144).

The Mean

Using the Gaussian window defined in (142), and integrating over ω and $\tilde{\omega}$ in (144), we obtain the following expression of the mean of the correlations

$$\mathbb{E}\left[\mathcal{C}(t,\tilde{t};\vec{\mathbf{x}},\vec{\mathbf{x}}')\right] \approx \frac{\sqrt{2\pi}\, e^{-\frac{|\mathbf{x}-\mathbf{x}'|^2}{2X^2(\omega_o)}}}{(4\pi L)^2\sqrt{4B^2/\Omega^2+1}} \\ e^{-i\omega_o\left[\tilde{\tau}_o(\vec{\mathbf{x}},\vec{\mathbf{x}}',\vec{\mathbf{y}})-\tilde{t}\right]-\frac{B^2\left[\tilde{\tau}_o(\vec{\mathbf{x}},\vec{\mathbf{x}}',\vec{\mathbf{y}})-\tilde{t}\right]^2}{4}-\frac{B^2\left[\overline{\tau}_o(\vec{\mathbf{x}},\vec{\mathbf{x}}',\vec{\mathbf{y}})-t\right]^2}{(4B^2/\Omega^2+1)}}. \tag{145}$$

They are indeed peaked at $t = \overline{\tau}_o(\vec{\mathbf{x}},\vec{\mathbf{x}}',\vec{\mathbf{y}})$ and $\tilde{t} = \tilde{\tau}_o(\vec{\mathbf{x}},\vec{\mathbf{x}}',\vec{\mathbf{y}})$. The width of the peak in $\tilde{t}$ is inherited from the $O(1/B)$ support of the pulse. The width of the peak in t is $O\left(\sqrt{1/B^2+4/\Omega^2}\right)$. It is similar to the width of the pulse $f(t)$ when the bandwidth satisfies $B \lesssim \Omega$, but it is much broader than $f(t)$ for larger bandwidths $B \gg \Omega$. The peak value of the mean is

$$\max_{t,\tilde{t}} \mathbb{E}\left[\mathcal{C}(t,\tilde{t};\vec{\mathbf{x}},\vec{\mathbf{x}}')\right] = \mathbb{E}\left[\mathcal{C}(\overline{\tau}_o(\vec{\mathbf{x}},\vec{\mathbf{x}}',\vec{\mathbf{y}}),\tilde{\tau}_o(\vec{\mathbf{x}},\vec{\mathbf{x}}',\vec{\mathbf{y}});\vec{\mathbf{x}},\vec{\mathbf{x}}')\right] \\ \approx \frac{\sqrt{2\pi}\, e^{-\frac{|\mathbf{x}-\mathbf{x}'|^2}{2X^2(\omega_o)}}}{(4\pi L)^2\sqrt{4B^2/\Omega^2+1}}. \tag{146}$$

It decays exponentially with the cross-range offset, which is why we restrict it with the window Ψ in the CINT function (140). This is essential for improving the SNR of the imaging function, as we show next.

The SNR

We assess the statistical stability of the peak of the cross-correlations by estimating the SNR, which is the ratio of (146) and the standard deviation of $\mathcal{C}(\overline{\tau}_o,\tilde{\tau}_o;\vec{\mathbf{x}},\vec{\mathbf{x}}')$. We need the second moment

$$\mathbb{E}\left[\left|\mathcal{C}(\overline{\tau}_o(\vec{\mathbf{x}},\vec{\mathbf{x}}',\vec{\mathbf{y}}),\tilde{\tau}_o(\vec{\mathbf{x}},\vec{\mathbf{x}}',\vec{\mathbf{y}});\vec{\mathbf{x}},\vec{\mathbf{x}}')\right|^2\right] \\ = \frac{1}{(32\pi^3L^2)^2B^4}\iint_{-\infty}^{\infty} d\omega_1\, d\omega_2\, e^{-\frac{(\omega_1-\omega_o)^2}{B^2}-\frac{(\omega_2-\omega_o)^2}{B^2}}\iint_{-\infty}^{\infty} d\tilde{\omega}_1\, d\tilde{\omega}_2 \\ \times \hat{\Phi}\left(\frac{\tilde{\omega}_1}{\Omega_\Phi}\right)\overline{\hat{\Phi}}\left(\frac{\tilde{\omega}_2}{\Omega_\Phi}\right)e^{-\frac{\tilde{\omega}_1^2+\tilde{\omega}_2^2}{4B^2}}\mathbb{E}\left[e^{i\left(\omega_1-\omega_2+\frac{\tilde{\omega}_1-\tilde{\omega}_2}{2}\right)\nu_\tau(\vec{\mathbf{x}},\vec{\mathbf{y}})-i\left(\omega_1-\omega_2-\frac{\tilde{\omega}_1-\tilde{\omega}_2}{2}\right)\nu_\tau(\vec{\mathbf{x}}',\vec{\mathbf{y}})}\right], \tag{147}$$

where we let the support of the windows be Ω_Φ, not necessarily the same as Ω, in order to demonstrate how it affects the SNR. The peak value of the mean for this choice is a slight modification of (146)

$$\mathbb{E}\left[\mathcal{C}(\overline{\tau}_o(\vec{\mathbf{x}},\vec{\mathbf{x}}',\vec{\mathbf{y}}),\tilde{\tau}_o(\vec{\mathbf{x}},\vec{\mathbf{x}}',\vec{\mathbf{y}});\vec{\mathbf{x}},\vec{\mathbf{x}}')\right] \approx \frac{\sqrt{2\pi}\, e^{-\frac{|\mathbf{x}-\mathbf{x}'|^2}{2X^2(\omega_o)}}}{(4\pi L)^2\sqrt{2B^2/\Omega^2+2B^2/\Omega_\Phi^2+1}}\,, \tag{148}$$

and the second moment follows from the calculation in Appendix 2

$$\mathbb{E}\left[\left|\mathcal{C}(\overline{\tau}_o(\vec{\mathbf{x}},\vec{\mathbf{x}}',\vec{\mathbf{y}}),\tilde{\tau}_o(\vec{\mathbf{x}},\vec{\mathbf{x}}',\vec{\mathbf{y}});\vec{\mathbf{x}},\vec{\mathbf{x}}')\right|^2\right] \approx \frac{\left(\frac{2B^2}{\Omega_\Phi^2}+1\right)^{-\frac{1}{2}}}{128\pi^3 L^4}$$
$$\left[\frac{2B^2}{\Omega^2}\left(1-\int_0^1 dt\, e^{-\frac{t^2|\mathbf{x}-\mathbf{x}'|^2}{2\ell^2}}\right)+1\right]^{-\frac{1}{2}}\times$$
$$\left[\frac{2B^2}{\Omega^2}\left(1-\int_0^1 dt\, e^{-\frac{t^2|\mathbf{x}-\mathbf{x}'|^2}{2\ell^2}}\right)+\frac{2B^2}{\Omega_\Phi^2}+1\right]^{-\frac{1}{2}}. \tag{149}$$

We distinguish two regimes:

1. When $|\mathbf{x}-\mathbf{x}'| \gg X(\omega_o)$ the mean (148) and therefore the SNR are exponentially small. The cross-correlations have large random fluctuations in this regime and in CINT we make sure that we do not use them by restricting $|\mathbf{x}-\mathbf{x}'| \lesssim X(\omega_o)$ with the window function Ψ.
2. When $|\mathbf{x}-\mathbf{x}'| \lesssim X(\omega_o)$, we can expand the exponential in (149) in series of $|\mathbf{x}-\mathbf{x}'|/\ell \ll 1$ to obtain

$$\frac{2B^2}{\Omega^2}\left(1-\int_0^1 dt\, e^{-\frac{t^2|\mathbf{x}-\mathbf{x}'|^2}{2\ell^2}}\right) \approx \frac{B^2}{\omega_o^2}\frac{|\mathbf{x}-\mathbf{x}'|^2}{X^2(\omega_o)} \ll 1\,,$$

and simplify (149) as

$$\mathbb{E}\left[\left|\mathcal{C}(\overline{\tau}_o(\vec{\mathbf{x}},\vec{\mathbf{x}}',\vec{\mathbf{y}}),\tilde{\tau}_o(\vec{\mathbf{x}},\vec{\mathbf{x}}',\vec{\mathbf{y}});\vec{\mathbf{x}},\vec{\mathbf{x}}')\right|^2\right] \approx \frac{\left(\frac{2B^2}{\Omega_\Phi^2}+1\right)^{-1}}{128\pi^3 L^4}\,. \tag{150}$$

The SNR is bounded by its value at $\mathbf{x}=\mathbf{x}'$, in which case Eqs. (148) and (150) give

$$\mathrm{SNR} \lesssim \frac{\left|\mathbb{E}\left[\mathcal{C}(\overline{\tau}_o(\vec{\mathbf{x}},\vec{\mathbf{x}},\vec{\mathbf{y}}),\tilde{\tau}_o(\vec{\mathbf{x}},\vec{\mathbf{x}},\vec{\mathbf{y}});\vec{\mathbf{x}},\vec{\mathbf{x}})\right]\right|}{\sqrt{\mathrm{Var}\left[\mathcal{C}(\overline{\tau}_o(\vec{\mathbf{x}},\vec{\mathbf{x}},\vec{\mathbf{y}}),\tilde{\tau}_o(\vec{\mathbf{x}},\vec{\mathbf{x}},\vec{\mathbf{y}});\vec{\mathbf{x}},\vec{\mathbf{x}})\right]}} \approx \sqrt{\frac{\Omega^2}{\Omega_\Phi^2}+\frac{\Omega^2}{2B^2}}\,. \tag{151}$$

The subtraction of the random phases in the local cross-correlations (143) at nearby receivers reduces substantially the random fluctuations and the SNR is no

longer exponentially small. However, the SNR may still not be large enough and it decreases when we increase Ω_Φ, as stated before. A small Ω_Φ gives a larger SNR but it also reduces the resolution as discussed previously, so $\Omega_\Phi \approx \Omega$ is a good compromise. Equation (151) shows that the stronger the random medium effects (i.e., the smaller Ω) the lower the SNR. The local cross-correlations are less sensitive to the realization of the random medium than the data themselves, but still they are not statistically stable. They have random fluctuations that are of the same order as the mean. These fluctuations are averaged out in the CINT imaging function (140) by the summation over the sensors, if the aperture is large enough. This is how we can achieve a statistically stable image, as we show in the next section.

Resolution Analysis of the CINT Imaging Function

We calculate first the mean of the imaging function to estimate the resolution of the expected focusing. The we estimate the SNR and state the conditions under which CINT is statistically stable.

The Mean Point Spread Function

In the case of a point source at $\vec{\mathbf{y}}$, the expectation of the imaging function follows directly from (140) and (145). Using the Gaussian window Ψ defined in (142) and the approximation (126) of the differences of travel times, we obtain

$$\mathbb{E}\left[\mathcal{J}(\vec{\mathbf{y}}^s;\vec{\mathbf{y}})\right] \approx \frac{\sqrt{2\pi}N^2}{(4\pi L)^2|\mathcal{A}|^2\sqrt{4B^2/\Omega^2+1}} \iint_{\mathcal{A}} d\mathbf{x}d\mathbf{x}' e^{-\frac{|\mathbf{x}-\mathbf{x}'|^2}{X^2(\omega_o)}+ik_o\frac{(\mathbf{y}^s-\mathbf{y})\cdot(\mathbf{x}-\mathbf{x}')}{L}-\frac{B^2}{4c_o^2}\left[\frac{(\mathbf{y}^s-\mathbf{y})\cdot(\mathbf{x}-\mathbf{x}')}{L}\right]^2} \times e^{-\frac{B^2}{c_o^2(4B^2/\Omega^2+1)}\left[\eta^s-\eta+\frac{(\mathbf{y}^s-\mathbf{y})}{L}\cdot\left(\frac{\mathbf{x}+\mathbf{x}'}{2}\right)\right]^2}. \tag{152}$$

Here we used the same notation as in the previous section for the point spread function $\mathcal{J}(\vec{\mathbf{y}}^s;\vec{\mathbf{y}})$ to emphasize that the source is a point at $\vec{\mathbf{y}}$.

It is convenient to change variables in the integral in (152) to centered and difference offsets

$$\overline{\mathbf{x}} = \frac{\mathbf{x}+\mathbf{x}'}{2}, \qquad \tilde{\mathbf{x}} = \mathbf{x}-\mathbf{x}',$$

and since only $|\tilde{\mathbf{x}}| \lesssim X(\omega_o) \ll \ell \lesssim a$ is in the essential support of the Gaussian, we can approximate the right hand side in (152) by extending the integral over $\tilde{\mathbf{x}}$ to the whole $\mathbb{R}^2$. Moreover, we note that under our scaling assumptions

$$\frac{B}{c_o}\left|\frac{(\mathbf{y}^s-\mathbf{y})\cdot(\mathbf{x}-\mathbf{x}')}{L}\right| \lesssim \frac{B}{\omega_o}\frac{k_o|\mathbf{y}^s-\mathbf{y}|X(\omega_o)}{L} \ll 1,$$

so we can neglect the third term in the exponential in (152). This is for search points $\vec{\mathbf{y}}^s$ satisfying

$$1 \lesssim \frac{|\mathbf{y}^s - \mathbf{y}|}{\lambda_o L / X(\omega_o)} \ll \frac{\omega_o}{B},$$

where $\lambda_o L / X(\omega_o)$ turns out to be the cross-range resolution limit, as shown below. We obtain after integrating in $\tilde{\mathbf{x}}$ that

$$\mathbb{E}\left[\mathcal{J}(\vec{\mathbf{y}}^s; \vec{\mathbf{y}})\right] \approx \frac{\sqrt{2/\pi} N^2 X^2(\omega_o)}{16 L^2 |\mathcal{A}| \sqrt{4B^2/\Omega^2 + 1}} e^{-\frac{1}{4}\left[\frac{k_o X(\omega_o)|\mathbf{y}^s - \mathbf{y}|}{L}\right]^2} \int_{\mathcal{A}} \frac{d\bar{\mathbf{x}}}{|\mathcal{A}|} e^{-\frac{B^2}{c_o^2(4B^2/\Omega^2+1)}\left[\eta^s - \eta + \frac{(\mathbf{y}^s - \mathbf{y})\cdot\bar{\mathbf{x}}}{L}\right]^2}, \tag{153}$$

and note that the right-hand side peaks at $\vec{\mathbf{y}}^s = \vec{\mathbf{y}}$, as it should.

Range Resolution When we evaluate (153) at $\mathbf{y}^s = \mathbf{y}$, we obtain

$$\mathbb{E}\left[\mathcal{J}(\vec{\mathbf{y}}^s = (\mathbf{y}, \eta^s); \vec{\mathbf{y}})\right] \approx \frac{\sqrt{2/\pi} N^2 X^2(\omega_o)}{16 L^2 |\mathcal{A}| \sqrt{4B^2/\Omega^2 + 1}} e^{-\frac{B^2(\eta^s - \eta)^2}{c_o^2(4B^2/\Omega^2+1)}}. \tag{154}$$

The range resolution is given by

$$|\eta^s - \eta| \lesssim \frac{c_o}{B}\sqrt{4B^2/\Omega^2 + 1}. \tag{155}$$

It is worse than in the homogeneous medium because of the factor $\sqrt{4B^2/\Omega^2 + 1}$. This factor is large when the decoherence frequency satisfies $\Omega \ll B$ and reduces the range resolution to $O(c_o/\Omega)$.

Cross-Range Resolution Evaluating (153) at $\eta^s = \eta$, we obtain

$$\mathbb{E}\left[\mathcal{J}(\vec{\mathbf{y}}^s; \vec{\mathbf{y}})\right] \approx \frac{\sqrt{\pi/2} N^2 X^2(\omega_o)}{32 L^2 |\mathcal{A}| \sqrt{4B^2/\Omega^2 + 1}} e^{-\frac{1}{4}\left[\frac{k_o X(\omega_o)|\mathbf{y}^s - \mathbf{y}|}{L}\right]^2} \prod_{j=1}^{2} \frac{\operatorname{erf}\left[\frac{aB}{2Lc_o\sqrt{4B^2/\Omega^2+1}}(y_j^s - y_j)\right]}{\frac{aB}{2Lc_o\sqrt{4B^2/\Omega^2+1}}(y_j^s - y_j)}.$$

Both factors are peaked at $\mathbf{y}^s = \mathbf{y}$, so we estimate the resolution by the width of the tighter peak

$$|\mathbf{y}^s - \mathbf{y}| \lesssim \min\left\{ \frac{\lambda_o L}{X(\omega_o)}, \frac{\lambda_o L}{a}\sqrt{\frac{4\omega_o^2}{\Omega^2} + \frac{\omega_o^2}{B^2}} \right\} . \tag{156}$$

This is worse than the resolution $\lambda_o L/a$ in homogeneous media because in our scaling $X(\omega_o) \ll a$ and $\sqrt{\frac{4\omega_o^2}{\Omega^2} + \frac{\omega_o^2}{B^2}} \gg 1$.

The SNR

To calculate the SNR, we need the variance of $\mathcal{J}(\vec{\mathbf{y}}^s; \vec{\mathbf{y}})$ and therefore the fourth-order moments of the Green's function. Their calculation is straightforward, due to the Gaussianity of ν_τ, but laborious. We state directly the result which distinguishes between two cases:

1. When $|\Delta\mathbf{x}| = |\mathbf{x} - \mathbf{x}'| \gg \ell$, the fourth moments

$$\mathcal{M}_4(\omega, \omega', \tilde{\omega}, \tilde{\omega}', \mathbf{x}, \mathbf{x}', \tilde{\mathbf{x}}, \tilde{\mathbf{x}}') = \mathbb{E}\Bigg[e^{i\left(\omega+\frac{\tilde{\omega}}{2}\right)\nu_\tau\left(\vec{\mathbf{x}}+\frac{\vec{\tilde{\mathbf{x}}}}{2},\vec{\mathbf{y}}\right) - i\left(\omega-\frac{\tilde{\omega}}{2}\right)\nu_\tau\left(\vec{\mathbf{x}}-\frac{\vec{\tilde{\mathbf{x}}}}{2},\vec{\mathbf{y}}\right)} \times e^{-i\left(\omega'+\frac{\tilde{\omega}'}{2}\right)\nu_\tau\left(\vec{\mathbf{x}}'+\frac{\vec{\tilde{\mathbf{x}}}'}{2},\vec{\mathbf{y}}\right) + i\left(\omega'-\frac{\tilde{\omega}'}{2}\right)\nu_\tau\left(\vec{\mathbf{x}}'-\frac{\vec{\tilde{\mathbf{x}}}'}{2},\vec{\mathbf{y}}\right)} \Bigg]$$

   factorize in the product of second moments

$$\mathcal{M}_4(\omega, \omega', \tilde{\omega}, \tilde{\omega}', \mathbf{x}, \mathbf{x}', \tilde{\mathbf{x}}, \tilde{\mathbf{x}}') \approx \mathcal{M}_2(\omega, \tilde{\omega}, \mathbf{x}, \tilde{\mathbf{x}})\overline{\mathcal{M}_2(\omega', \tilde{\omega}', \mathbf{x}', \tilde{\mathbf{x}}')} ,$$

   where

$$\mathcal{M}_2(\omega, \tilde{\omega}, \mathbf{x}, \tilde{\mathbf{x}}) = \mathbb{E}\left[e^{i\left(\omega+\frac{\tilde{\omega}}{2}\right)\nu_\tau\left(\vec{\mathbf{x}}+\frac{\vec{\tilde{\mathbf{x}}}}{2},\vec{\mathbf{y}}\right) - i\left(\omega-\frac{\tilde{\omega}}{2}\right)\nu_\tau\left(\vec{\mathbf{x}}-\frac{\vec{\tilde{\mathbf{x}}}}{2},\vec{\mathbf{y}}\right)} \right] \approx e^{-\frac{\tilde{\omega}^2}{2\Omega^2} - \frac{|\tilde{\mathbf{x}}|^2}{2X^2(\omega_o)}} ,$$

   and $\vec{\mathbf{x}} = (\mathbf{x}, L)$, $\vec{\tilde{\mathbf{x}}} = (\tilde{\mathbf{x}}, 0)$. The same notation applies for the prime variables.
2. When $|\Delta\mathbf{x}| \lesssim \ell$, we have the expression

$$\mathcal{M}_4(\omega, \omega', \tilde{\omega}, \tilde{\omega}', \mathbf{x}, \mathbf{x}', \tilde{\mathbf{x}}, \tilde{\mathbf{x}}') \approx \mathcal{M}_2(\omega, \tilde{\omega}, \mathbf{x}, \tilde{\mathbf{x}})\overline{\mathcal{M}_2(\omega', \tilde{\omega}', \mathbf{x}', \tilde{\mathbf{x}}')}$$
$$\times \exp\left\{ \int_0^1 dt\, e^{-\frac{t^2|\Delta\mathbf{x}|^2}{2\ell^2}} \left[\frac{\tilde{\omega}\tilde{\omega}'}{\Omega^2} + 3\frac{t^2\tilde{\mathbf{x}}\cdot\tilde{\mathbf{x}}'}{X^2(\omega_o)}\left(1 - \frac{t^2|\Delta\mathbf{x}|^2}{2\ell^2}\right) \right.\right.$$
$$\left.\left. -\sqrt{3}t^2\frac{\Delta\mathbf{x}}{\ell}\cdot\left(\frac{\tilde{\omega}'\tilde{\mathbf{x}}}{\Omega X(\omega_o)} - \frac{\tilde{\omega}\tilde{\mathbf{x}}'}{\Omega X(\omega_o)}\right)\right] \right\} .$$

   We no longer have the factorization of the moments, and the exponential factor is bounded above and below by constants of order one.

The factorization of the moments for $|\Delta \mathbf{x}| \gg \ell$ means that only the set $\{\mathbf{x}, \mathbf{x}' \in \mathcal{A}, \quad |\mathbf{x}-\mathbf{x}'| \lesssim \ell\}$ contributes to the calculation of the variance of $\mathcal{J}(\vec{\mathbf{y}}^s)$. Explicitly, we have

$$\mathrm{Var}\left[\mathcal{J}(\vec{\mathbf{y}};\vec{\mathbf{y}})\right] \approx \frac{N^4}{64|\mathcal{A}|^4\pi^5 L^4 B^4} \iint_{-\infty}^{\infty} d\omega d\omega' \iint_{-\infty}^{\infty} d\tilde{\omega} d\tilde{\omega}' \iint_{\mathcal{A}} d\mathbf{x} d\mathbf{x}' 1_{[0,\ell]}$$
$$(|\mathbf{x}-\mathbf{x}'|) \iint_{\mathbb{R}^{\not\perp}} d\tilde{\mathbf{x}} d\tilde{\mathbf{x}}' \times e^{-\frac{(\omega-\omega_o)^2+(\omega'-\omega_o)^2}{B^2} - (\tilde{\omega}^2+\tilde{\omega}'^2)\left(\frac{1}{\Omega^2}+\frac{1}{4B^2}\right) - \frac{|\tilde{\mathbf{x}}|^2+|\tilde{\mathbf{x}}'|^2}{2X^2(\omega_o)}}$$
$$\times \left\{ e^{\int_0^1 dt\, e^{-\frac{t^2|\Delta\mathbf{x}|^2}{2\ell^2}} \left[\frac{\tilde{\omega}\tilde{\omega}'}{\Omega^2} + 3\frac{t^2\tilde{\mathbf{x}}\cdot\tilde{\mathbf{x}}'}{X^2(\omega_o)}\left(1-\frac{t^2|\Delta\mathbf{x}|^2}{2\ell^2}\right) - \sqrt{3}t^2\frac{\Delta\mathbf{x}}{\ell}\cdot\left(\frac{\tilde{\omega}'\tilde{\mathbf{x}}}{\Omega X(\omega_o)} - \frac{\tilde{\omega}\tilde{\mathbf{x}}'}{\Omega X(\omega_o)}\right)\right]} - 1 \right\}.$$

This is a complicated expression but we can estimate it by replacing the last factor by an order one constant. We obtain that

$$\mathrm{Var}\left[\mathcal{J}(\vec{\mathbf{y}};\vec{\mathbf{y}})\right] \sim \frac{\ell^2}{|\mathcal{A}|}\left|\mathbb{E}\left[\mathcal{J}(\vec{\mathbf{y}};\vec{\mathbf{y}})\right]\right|^2, \tag{157}$$

where the symbol $\sim$ means approximate, up to a multiplicative constant of order one. The SNR at the peak is given by

$$\mathrm{SNR}\left[\mathcal{J}(\vec{\mathbf{y}};\vec{\mathbf{y}})\right] \sim \frac{a}{\ell}, \tag{158}$$

where we used that $|\mathcal{A}| = a^2$. Thus, the CINT point spread function $\mathcal{J}(\vec{\mathbf{y}}^s;\vec{\mathbf{y}})$ has a robust (statistically stable) focusing at the location $\vec{\mathbf{y}}$ of the source only if the array has aperture $a \gg \ell$. This is a significant improvement over the result in Sect. 7, where for the migration method to be stable we needed that a/ℓ be much larger than a huge number, not one like above.

Remark 4. Our analysis above is restricted to the CINT point spread function. It extends easily to the case of a distributed source density. We do not do it here because the calculations are long and the result does not bring any significant insight to the problem.

CINT Images for Passive Arrays as the Smoothed Wigner Transform

The Wigner transform, also called the Wigner distribution, is a classic tool for studying high frequency limits of the wave equation. It is very useful in imaging because it allows us to extract the important phase information from the measurements at the array: the travel times and the direction of arrival of the waves from the unknown source location.

The Wigner transform of the wave field $\hat{p}(\omega,(\mathbf{x},L))$ evaluated in the plane of the array, at range L from the source, is defined by

$$\mathcal{W}(\omega,\mathbf{x};t,\mathbf{K}) = \int_{\mathbb{R}^2} \frac{d\tilde{\mathbf{x}}}{(2\pi)^2} \int_{-\infty}^{\infty} \frac{d\tilde{\omega}}{2\pi}\, \hat{p}\left(\omega+\frac{\tilde{\omega}}{2},\left(\mathbf{x}+\frac{\tilde{\mathbf{x}}}{2},L\right)\right) \overline{\hat{p}}\left(\omega-\frac{\tilde{\omega}}{2},\left(\mathbf{x}-\frac{\tilde{\mathbf{x}}}{2},L\right)\right) e^{-i\omega\mathbf{K}\cdot\tilde{\mathbf{x}}-i\tilde{\omega}t}. \tag{159}$$

We cannot compute it in practice because the measurements of $\hat{p}$ are limited to the aperture, but we shall see in this section how $\mathcal{W}$ is related to the CINT imaging function. Note that $\mathcal{W}$ is real valued and its integral over the last two arguments gives the intensity of the wave field

$$\int_{\mathbb{R}^2} d\mathbf{K}\,\omega^2 \int_{-\infty}^{\infty} dt\, \mathcal{W}(\omega,\mathbf{x};t,\mathbf{K}) = \left|\hat{p}\left(\omega,(\mathbf{x},L)\right)\right|^2. \tag{160}$$

Thus, we may think of W as the "energy density" of the waves resolved over the phase space variables t and $\mathbf{K}$, although $\mathcal{W}$ is not positive in general. The variable t is dual to the frequency offset $\tilde{\omega}$, so it has units of time. The variable $\mathbf{K}$ is dual to the cross-range offset, and is called the slowness vector because its units are time over length, like $1/c_o$.

We use the paraxial approximation (112) of the travel time and assumption (118) to rewrite the expression (139) of the CINT imaging function $\mathcal{J}(\vec{\mathbf{y}}^s)$ evaluated at $\vec{\mathbf{y}}^s = (\mathbf{y}^s,\eta^s)$ as

$$\mathcal{J}(\vec{\mathbf{y}}^s) \approx \frac{N^2}{(2\pi)^2|\mathcal{A}|^2} \int_{-\infty}^{\infty} d\omega \int_{-\infty}^{\infty} d\tilde{\omega}\, \hat{\Phi}\left(\frac{\tilde{\omega}}{\Omega}\right) \int_{\mathcal{A}} d\mathbf{x} \int_{\mathbb{R}^2} d\tilde{\mathbf{x}}\, \Psi\left[\frac{\tilde{\mathbf{x}}}{X(\omega_o)}\right] \times \hat{p}\left(\omega+\frac{\tilde{\omega}}{2},\left(\mathbf{x}+\frac{\tilde{\mathbf{x}}}{2},L\right)\right) \overline{\hat{p}}\left(\omega-\frac{\tilde{\omega}}{2},\left(\mathbf{x}-\frac{\tilde{\mathbf{x}}}{2},L\right)\right) e^{-i\tilde{\omega}\tau_o\left((\mathbf{x},L),\vec{\mathbf{y}}^s\right)-i\frac{\omega}{c_o}\frac{(\mathbf{x}-\mathbf{y}^s)\cdot\tilde{\mathbf{x}}}{L}}. \tag{161}$$

The cross-range variables $\mathbf{x}$ and $\tilde{\mathbf{x}}$ should take values in the set defined by the constraint $\mathbf{x}\pm\tilde{\mathbf{x}}/2\in\mathcal{A}$, so that the wave field is evaluated in the aperture of the array. Nevertheless, since the window Ψ restricts $|\tilde{\mathbf{x}}|$ to the decorrelation length $X(\omega_o)$, which is much smaller than the aperture a, we can approximate $\mathcal{J}(\vec{\mathbf{y}}^s)$ by extending the integral over $\tilde{\mathbf{x}}$ to the whole plane and letting $\mathbf{x}$ vary in $\mathcal{A}$.

The relation between $\mathcal{J}(\vec{\mathbf{y}}^s)$ and $\mathcal{W}$ follows from (161), after inverting the transform in (159),

$$\mathcal{J}(\vec{\mathbf{y}}^s) \approx \frac{N^2\Omega X^2(\omega_o)}{2\pi|\mathcal{A}|^2} \int_{-\infty}^{\infty} d\omega \int_{\mathcal{A}} d\mathbf{x} \int_{-\infty}^{\infty} dt \int_{\mathbb{R}^2} d\mathbf{K}\,\omega^2 \mathcal{W}(\omega,\mathbf{x};t,\mathbf{K}) \times \Phi\left[\Omega\left(\tau_o((\mathbf{x},L),\vec{\mathbf{y}}^s)-t\right)\right] \hat{\Psi}\left[\omega_o X(\omega_o)\left(\mathbf{K}-\frac{\mathbf{x}-\mathbf{y}^s}{c_o L}\right)\right]. \tag{162}$$

Thus $\mathcal{J}(\vec{\mathbf{y}}^s)$ is given by the Wigner transform smoothed over all its arguments. The smoothing is by the integration over the bandwidth and the aperture and by the convolution with the windows. The time window Φ has support of order $1/\Omega$, and the convolution is evaluated at the travel time from the search point $\vec{\mathbf{y}}^s$ to the receiver location $\vec{\mathbf{x}} = (\mathbf{x}, L)$ in the array. The window $\hat{\Psi}$ over the slowness vectors has support of order $\omega_o X(\omega_o) = \sqrt{3}\Omega\ell$, and the convolution is evaluated at the slowness vector $(\mathbf{x}-\mathbf{y}^s)/(c_o L)$, which is approximately the cross-range gradient of the travel time $\nabla_{\mathbf{x}}\tau_o(\vec{\mathbf{x}}, \vec{\mathbf{y}}^s)$.

Remark 5. It turns out that the Wigner transform is weakly self-averaging for a wide range of wave scattering regimes in random media. That is to say, although $\mathcal{W}(\omega, \mathbf{x}; t, \mathbf{K})$ is random when evaluated pointwise, it becomes deterministic when integrated against some test function $\varphi(\omega, \mathbf{x}; t, \mathbf{K})$, as in (162). The weak self-averaging property of the Wigner transform is the basis of the proofs of statistical stability of the time reversal process [3, 27] and of CINT imaging in random media [9]. It is an asymptotic result obtained by taking various limits, depending on the scaling regime. The smoothed Wigner transform is not deterministic in practice because the limit is never realized, but its random fluctuations at the peaks are small, i.e., the SNR is large.

Calculation of the Wigner Transform

Here we calculate the Wigner transform in order to show explicitly the role of the smoothing by the windows Φ and Ψ used in the local cross-correlations of the array data. Again, we assume for simplicity a point source at $\vec{\mathbf{y}}$.

Using the model (137) of the array data with Green's function (122) in the definition (159) of the Wigner transform, we obtain

$$\mathcal{W}(\omega, \mathbf{x}; t, K) \approx \frac{1}{(4\pi L)^2(2\pi)^3} \int_{\mathbb{R}^2} d\tilde{\mathbf{x}} \int_{-\infty}^{\infty} d\tilde{\omega}\, \hat{f}\left(\omega + \frac{\tilde{\omega}}{2}\right) \overline{\hat{f}}\left(\omega - \frac{\tilde{\omega}}{2}\right)$$
$$\times e^{i\omega\left[\tau\left(\left(\mathbf{x}+\frac{\tilde{\mathbf{x}}}{2},L\right),\vec{\mathbf{y}}\right)-\tau\left(\left(\mathbf{x}-\frac{\tilde{\mathbf{x}}}{2},L\right),\vec{\mathbf{y}}\right)-\mathbf{K}\cdot\tilde{\mathbf{x}}\right]+i\frac{\tilde{\omega}}{2}\left[\tau\left(\left(\mathbf{x}+\frac{\tilde{\mathbf{x}}}{2},L\right),\vec{\mathbf{y}}\right)+\tau\left(\left(\mathbf{x}-\frac{\tilde{\mathbf{x}}}{2},L\right),\vec{\mathbf{y}}\right)-2t\right]}. \tag{163}$$

Here τ is the random travel time

$$\tau(\vec{\mathbf{x}}, \vec{\mathbf{y}}) = \tau_o(\vec{\mathbf{x}}, \vec{\mathbf{y}}) + \nu_\tau(\vec{\mathbf{x}}, \vec{\mathbf{y}}),$$

and assuming that the pulse $\hat{f}$ is modeled by the Gaussian (105), we can integrate over $\tilde{\omega}$ in (163)

$$\mathcal{W}(\omega, \mathbf{x}; t, K) \approx \frac{e^{-(\omega-\omega_o)^2/B^2}}{(4\pi L)^2 2\pi^{3/2} B} \int_{\mathbb{R}^2} d\tilde{\mathbf{x}}\, e^{-\frac{B^2}{4}\left[\tau\left(\left(\mathbf{x}+\frac{\tilde{\mathbf{x}}}{2},L\right),\vec{\mathbf{y}}\right)+\tau\left(\left(\mathbf{x}-\frac{\tilde{\mathbf{x}}}{2},L\right),\vec{\mathbf{y}}\right)-2t\right]^2}$$
$$\times e^{i\omega\left[\tau\left(\left(\mathbf{x}+\frac{\tilde{\mathbf{x}}}{2},L\right),\vec{\mathbf{y}}\right)-\tau\left(\left(\mathbf{x}-\frac{\tilde{\mathbf{x}}}{2},L\right),\vec{\mathbf{y}}\right)-\mathbf{K}\cdot\tilde{\mathbf{x}}\right]}. \tag{164}$$

It remains to evaluate the $\tilde{\mathbf{x}}$ integral which we cannot do explicitly in (164) because there is no restriction on the cross-range offset $\tilde{\mathbf{x}}$. However, when we smooth $\mathcal{W}$ over the slowness vectors, as we do in the convolution with the window $\hat{\Psi}$ in (164), we see that only small cross-range offsets contribute to the results. Explicitly, we have

$$\int_{\mathbb{R}^{\not\perp}} d\mathbf{K}' \hat{\Psi}\left[\omega_o X(\omega_o)(\mathbf{K}' - \mathbf{K})\right] \mathcal{W}(\omega, \mathbf{x}; t; \mathbf{K}') \approx \frac{2\sqrt{\pi} e^{-(\omega-\omega_o)^2/B^2}}{(4\pi L)^2 B[\omega_o X(\omega_o)]^2}$$
$$e^{-B^2\left[\tau\left((\mathbf{x},L),\vec{\mathbf{y}}\right)-t\right]^2} \times \int_{\mathbb{R}^2} d\tilde{\mathbf{x}}\, \Psi\left(\frac{\tilde{\mathbf{x}}}{X(\omega_o)}\right) e^{i\omega_o\left[\nabla_{\mathbf{x}}\tau\left(\left(\mathbf{x},L\right),\vec{\mathbf{y}}\right)-\mathbf{K}\right]\cdot\tilde{\mathbf{x}}}, \tag{165}$$

where we used our scaling assumptions to approximate

$$\frac{B}{2}\left[\tau\left((\mathbf{x}+\frac{\tilde{\mathbf{x}}}{2}, L), \vec{\mathbf{y}}\right) + \tau\left((\mathbf{x}-\frac{\tilde{\mathbf{x}}}{2}, L), \vec{\mathbf{y}}\right)\right] \approx B\tau\left((\mathbf{x}, L), \vec{\mathbf{y}}\right),$$
$$\omega\left[\tau\left((\mathbf{x}+\frac{\tilde{\mathbf{x}}}{2}, L), \vec{\mathbf{y}}\right) - \tau\left((\mathbf{x}-\frac{\tilde{\mathbf{x}}}{2}, L), \vec{\mathbf{y}}\right)\right] \approx \omega_o \nabla_{\mathbf{x}}\tau\left((\mathbf{x}, L), \vec{\mathbf{y}}\right),$$

for cross-range offsets in the support of Ψ, satisfying $|\tilde{\mathbf{x}}| \lesssim X(\omega_o)$. Now we can integrate over $\tilde{\mathbf{x}}$ in (165) and over ω

$$\int_{\mathbb{R}^{\not\perp}} d\mathbf{K}' \hat{\Psi}\left[\omega_o X(\omega_o)(\mathbf{K}' - \mathbf{K})\right] \int_{-\infty}^{\infty} d\omega\, \mathcal{W}(\omega, \mathbf{x}; t; \mathbf{K}')$$
$$\approx \frac{2\pi}{(4\pi L)^2\omega_o^2} e^{-B^2\left[\tau\left((\mathbf{x},L),\vec{\mathbf{y}}\right)-t\right]^2} \times \hat{\Psi}\left[\omega_o X(\omega_o)\left(\nabla_{\mathbf{x}}\tau\left((\mathbf{x}, L), \vec{\mathbf{y}}\right) - \mathbf{K}\right)\right]. \tag{166}$$

We conclude from (166) and the paraxial approximation (112) of the deterministic travel time τ_o that the Wigner transform smoothed over the slowness vector peaks at the random arguments

$$t = \tau_o\left((\mathbf{x}, L), \vec{\mathbf{y}}\right) + \nu_\tau\left((\mathbf{x}, L), \vec{\mathbf{y}}\right) \quad \text{and} \quad \mathbf{K} = \frac{\mathbf{x}-\mathbf{y}}{c_o L} + \nabla_{\mathbf{x}}\nu_\tau\left((\mathbf{x}, L), \vec{\mathbf{y}}\right). \tag{167}$$

The peak dances around the point $\left(\tau_o\left((\mathbf{x}, L), \vec{\mathbf{y}}\right), \frac{\mathbf{x}-\mathbf{y}}{c_o L}\right)$ in the $(t, \mathbf{K})$ phase space, as modeled by ν_τ in the t direction and by $\nabla_{\mathbf{x}}\nu_\tau$ in the plane of the slowness vectors $\mathbf{K}$. In the CINT imaging function (163) we evaluate (166) at the expected peak location $\left(\tau_o\left((\mathbf{x}, L), \vec{\mathbf{y}}^s\right), \frac{\mathbf{x}-\mathbf{y}^s}{c_o L}\right)$ supposing that the source is at $\vec{\mathbf{y}}^s$. The role of the window $\hat{\Psi}$ is to mitigate the peak dancing in $\mathbf{K}$. The standard deviation of the location of the peaks is quantified in our model as $O(1/(\omega_o X(\omega_o)))$, which is precisely the support of $\hat{\Psi}$ in (166). Thus, in CINT we are essentially taking the envelope of the random peaks in $\mathbf{K}$ to stabilize statistically the image, and thus increase the SNR.

The standard deviation of the peak location in t is $O(1/\Omega)$ in our model, and the peak dancing is significant in (166) when $B \gg \Omega$. The CINT imaging function

(163) is statistically stable because of the convolution with the window Φ of support $O(1/\Omega)$.

Remark 6. The results above reveal that on one hand the windows used in the calculation of the local cross-correlations are needed to stabilize the imaging process, but on the other hand they blur the peaks of $\mathcal{W}$ and therefore of the image. In practice it is unlikely that we would know the statistics of the medium fluctuations so we cannot calculate the decorrelation length $X(\omega_o)$ and the decoherence frequency Ω using the formulas (103) and (109). In fact these formulas may not even apply because the random travel time model may not be a good approximation of the net scattering effects of the medium. It is the form of the second moment formulas (109) which is general, and not the detailed definition of Ω and $X(\omega_o)$ that the CINT imaging methodology should use. The scales $X(\omega_o)$ and Ω should not be set based on a specific model, they should be estimated as part of the imaging process. This can be done in principle directly from the data, by calculating for example empirical correlations and estimating how they decay with the frequency and cross-range offsets. Alternatively, we may estimate $X(\omega_o)$ and Ω while we form the image, by exploiting the trade-off between the resolution (sharpness) of the image and its SNR. The latter approach is known as adaptive CINT. It uses a figure of merit of the quality of the image and estimates $X(\omega_o)$ and Ω by optimizing it during the image formation [8].

CINT Imaging with Active Arrays

The CINT imaging function for active arrays back-propagates the local cross-correlations of the array measurements $\hat{p}(\omega, \vec{\mathbf{x}}_r, \vec{\mathbf{x}}_s)$ to the search points $\vec{\mathbf{y}}^s$, using the roundtrip travel times

$$\tau_o(\vec{\mathbf{x}}_r, \vec{\mathbf{y}}^s, \vec{\mathbf{x}}_s) = \tau_o(\vec{\mathbf{x}}_r, \vec{\mathbf{y}}^s) + \tau_0(\vec{\mathbf{x}}_s, \vec{\mathbf{y}}^s). \tag{168}$$

Here $\vec{\mathbf{x}}_r = (\mathbf{x}_r, L)$ and $\vec{\mathbf{x}}_s = (\mathbf{x}_s, L)$ denote the location of the receivers and sources in the aperture $\mathcal{A}$, and the imaging function is given by

$$\begin{aligned}\mathcal{J}(\vec{\mathbf{y}}^s) = \sum_{r,r'=1}^{N_r} \sum_{s,s'=1}^{N_s} \int_{-\infty}^{\infty} \frac{d\omega}{2\pi} \Psi\left[\frac{\mathbf{x}_r - \mathbf{x}_{r'}}{X(\omega_o)}\right] \Psi\left[\frac{\mathbf{x}_s - \mathbf{x}_{s'}}{X(\omega_o)}\right] \int_{-\infty}^{\infty} \frac{d\tilde{\omega}}{2\pi} \hat{\Phi}\left(\frac{\tilde{\omega}}{\Omega}\right) \\ \hat{p}\left(\omega + \frac{\tilde{\omega}}{2}, \vec{\mathbf{x}}_r, \vec{\mathbf{x}}_s\right) \times \overline{\hat{p}}\left(\omega - \frac{\tilde{\omega}}{2}, \vec{\mathbf{x}}_{r'}, \vec{\mathbf{x}}_{s'}\right) \\ e^{-i\tilde{\omega}\left[\frac{\tau_o(\vec{\mathbf{x}}_r, \vec{\mathbf{y}}^s, \vec{\mathbf{x}}_s) + \tau_o(\vec{\mathbf{x}}_{r'}, \vec{\mathbf{y}}^s, \vec{\mathbf{x}}_{s'})}{2}\right] - i\omega[\tau_o(\vec{\mathbf{x}}_r, \vec{\mathbf{y}}^s, \vec{\mathbf{x}}_s) - \tau_o(\vec{\mathbf{x}}_{r'}, \vec{\mathbf{y}}^s, \vec{\mathbf{x}}_{s'})]}.\end{aligned} \tag{169}$$

The resolution analysis of (169) is more involved than that for passive arrays, even for the Born model of the data

$$\hat{p}(\omega,\vec{\mathbf{x}}_r,\vec{\mathbf{x}}_s) \approx \hat{f}(\omega)\int d\vec{\mathbf{y}}\,\rho(\vec{\mathbf{y}})\hat{G}(\omega,\vec{\mathbf{x}}_r,\vec{\mathbf{y}})\hat{G}(\omega,\vec{\mathbf{x}}_r,\vec{\mathbf{y}}^s)$$

$$\approx \frac{\hat{f}(\omega)}{(4\pi L)^2}\int d\vec{\mathbf{y}}\,\rho(\vec{\mathbf{y}})e^{i\omega\tau_o(\vec{\mathbf{x}}_r,\vec{\mathbf{y}},\vec{\mathbf{x}}_s)+i\omega\nu_\tau(\vec{\mathbf{x}}_r,\vec{\mathbf{y}})+i\omega\nu_\tau(\vec{\mathbf{x}}_s,\vec{\mathbf{y}})}, \tag{170}$$

because it requires higher moments of the random travel time perturbation ν_τ. We do not include it here because it does not bring any new insight. We only highlight the relation between the imaging function and the Wigner transform, in the case of a point-like reflector at $\vec{\mathbf{y}}$.

Using the model

$$\hat{p}(\omega,\vec{\mathbf{x}}_r,\vec{\mathbf{x}}_s) \approx \hat{f}(\omega)\hat{G}(\omega,\vec{\mathbf{x}}_r,\vec{\mathbf{y}})\hat{G}(\omega,\vec{\mathbf{x}}_s,\vec{\mathbf{y}}),$$

we can rewrite (169) in terms of the Wigner transform $W(\omega,\mathbf{x};t,\mathbf{K})$ of $\hat{f}(\omega)^{1/2}\hat{G}(\omega,\vec{\mathbf{x}},\vec{\mathbf{y}})$, which is like the square root of the data. Explicitly, we have that

$$W(\omega,\mathbf{x};t,\mathbf{K}) = \int_{\mathbb{R}^2}\frac{d\tilde{\mathbf{x}}}{(2\pi)^2}\int_{-\infty}^{\infty}\frac{d\tilde{\omega}}{2\pi}\,\hat{f}^{1/2}\left(\omega+\frac{\tilde{\omega}}{2}\right)\overline{\hat{f}^{1/2}}\left(\omega-\frac{\tilde{\omega}}{2}\right)$$
$$\hat{G}\left(\omega+\frac{\tilde{\omega}}{2},\left(\mathbf{x}+\frac{\tilde{\mathbf{x}}}{2},L\right)\right)\overline{\hat{G}}\left(\omega-\frac{\tilde{\omega}}{2},\left(\mathbf{x}-\frac{\tilde{\mathbf{x}}}{2},L\right)\right)e^{-i\omega\mathbf{K}\cdot\tilde{\mathbf{x}}-i\tilde{\omega}t},$$

and (169) becomes

$$\mathcal{J}(\vec{\mathbf{y}}^s) \approx \int_{-\infty}^{\infty}\frac{d\omega}{2\pi}\iint_{\mathbb{R}^2}d\mathbf{K}d\mathbf{K}'\iint_{-\infty}^{\infty}dtdt'\sum_{r,r'=1}^{N_r}\omega^2 W(\omega,\overline{\mathbf{x}}_r;t,\mathbf{K})\Psi\left[\frac{\tilde{\mathbf{x}}_r}{X(\omega_o)}\right]$$
$$\times\, e^{i\omega\left(\mathbf{K}-\frac{\overline{\mathbf{x}}_r-\mathbf{y}^s}{L}\right)\cdot\tilde{\mathbf{x}}_r}\times\sum_{s,s'=1}^{N_s}\omega^2 W(\omega,\overline{\mathbf{x}}_s;t',\mathbf{K}')\Psi\left[\frac{\tilde{\mathbf{x}}_s}{X(\omega_o)}\right]e^{i\omega\left(\mathbf{K}'-\frac{\overline{\mathbf{x}}_s-\mathbf{y}^s}{L}\right)\cdot\tilde{\mathbf{x}}_s}$$
$$\times\,\Omega\,\Phi\Big[\Omega\left(\tau_o\big((\mathbf{x}_r,L),\vec{\mathbf{y}}^s\big)-t+\tau_o\big((\mathbf{x}_s,L),\vec{\mathbf{y}}^s\big)-t'\right)\Big], \tag{171}$$

where we let

$$\overline{\mathbf{x}}_r = \frac{\mathbf{x}_r+\mathbf{x}_{r'}}{2} \quad\text{and}\quad \tilde{\mathbf{x}}_r = \mathbf{x}_r-\mathbf{x}_{r'},$$

integrated over $\tilde{\omega}$, and used the paraxial approximation (112) of the travel time and the assumption (118) on the decorrelation length.

The result (171) is similar to that for passive arrays. To see this more explicitly, let us suppose that $N_r = N_s$ and use the continuum aperture approximation to rewrite (171) as

$$\mathcal{J}(\vec{\mathbf{y}}^s) \approx \frac{N^4\Omega[\omega_o X(\omega_o)]^4}{2\pi|\mathcal{A}|^4}\int_{-\infty}^{\infty}\frac{d\omega}{2\pi}\iint_{-\infty}^{\infty}dt\,dt'$$
$$\iint_{\mathcal{A}} d\mathbf{x}d\mathbf{x}'\Phi\Big[\Omega\left(\tau_o\big((\mathbf{x},L),\vec{\mathbf{y}}^s\big)-t+\tau_o\big((\mathbf{x}',L),\vec{\mathbf{y}}^s\big)-t'\right)\Big]$$
$$\times\int_{\mathbb{R}^2} d\mathbf{K}\,W(\omega,\mathbf{x};t,\mathbf{K})\hat{\Psi}\left[\omega_o X(\omega_o)\left(\mathbf{K}-\frac{\mathbf{x}-\mathbf{y}^s}{L}\right)\right]$$
$$\times\int_{\mathbb{R}^2} d\mathbf{K}'\,W(\omega',\mathbf{x}';t',\mathbf{K}')\hat{\Psi}\left[\omega_o X(\omega_o)\left(\mathbf{K}'-\frac{\mathbf{x}'-\mathbf{y}^s}{L}\right)\right]. \tag{172}$$

The $\mathbf{K}$ and $\mathbf{K}'$ integrals are just like (165), except that now we have the square root of the pulse, so we can write directly the result from there,

$$\int_{\mathbb{R}^2} d\mathbf{K}\,W(\omega,\mathbf{x};t,\mathbf{K})\hat{\Psi}\left[\omega_o X(\omega_o)\left(\mathbf{K}-\frac{\mathbf{x}-\mathbf{y}^s}{L}\right)\right] \approx \frac{2e^{-\frac{(\omega-\omega_o)^2}{2B^2}}}{(4\pi L)^2\omega_o^2}e^{-2B^2\left[\tau((\mathbf{x},L),\vec{\mathbf{y}})-t\right]^2}$$
$$\times\,\hat{\Psi}\left[\omega_o X(\omega_o)\left(\nabla_{\mathbf{x}}\tau((\mathbf{x},L),\vec{\mathbf{y}})-\frac{\mathbf{x}-\mathbf{y}^s}{L}\right)\right].$$

We see the same random peak dancing as before, which is mitigated in the slowness plane by the window $\hat{\Psi}$ of width chosen appropriately as $1/(\omega_o X(\omega_o))$. The peak dancing in t is visible in the case $B \gg \Omega$, but it is mitigated in the imaging function by the convolution with the window Φ. This can be seen from (172), once we use the Gaussian window (142) and integrate in t and t'

$$\mathcal{J}(\vec{\mathbf{y}}^s) \approx \frac{N^4X^4(\omega_o)(2B^2/\Omega^2+1)^{-1/2}}{\sqrt{2\pi}(4\pi L)^4|\mathcal{A}|^4}$$
$$\iint_{\mathcal{A}} d\mathbf{x}d\mathbf{x}'\hat{\Psi}\left[\omega_o X(\omega_o)\left(\nabla_{\mathbf{x}}\tau((\mathbf{x},L),\vec{\mathbf{y}})-\frac{\mathbf{x}-\mathbf{y}^s}{L}\right)\right]$$
$$\times\,\hat{\Psi}\left[\omega_o X(\omega_o)\left(\nabla_{\mathbf{x}}\tau((\mathbf{x}',L),\vec{\mathbf{y}})-\frac{\mathbf{x}'-\mathbf{y}^s}{L}\right)\right]$$
$$\times\,e^{-\frac{B^2}{2B^2/\Omega^2+1}\left[\tau\left((\mathbf{x},L),\vec{\mathbf{y}}\right)+\tau\left((\mathbf{x}',L),\vec{\mathbf{y}}\right)-\tau_o\left((\mathbf{x},L),\vec{\mathbf{y}}^s\right)-\tau_o\left((\mathbf{x}',L),\vec{\mathbf{y}}\right)\right]^2}. \tag{173}$$

Numerical Simulations

We illustrate with numerical results the point spread function of CINT with active arrays. The array data $\hat{p}(\omega,\vec{\mathbf{x}}_r,\vec{\mathbf{x}}_s)$ is simulated with the random travel time approximation (78) of the Green's function in a medium with wave speed modeled by (73) for a constant $\bar{c} = c_o$. The standard deviation of the fluctuations of $c(\vec{\mathbf{x}})$ is either $\sigma = 0.04\,\%$ or $\sigma = 0.1\,\%$, and the correlation length is $\ell = 100\lambda_o$. The array has aperture $a = 4\ell$. It is centered at the cross-range location $\mathbf{y}$ of the reflector, and at range $L = 99\ell$ from it. There are $N = 101$ sensors in each direction. All

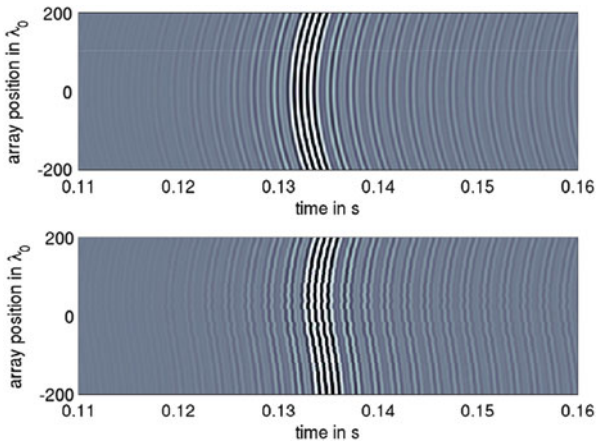


Fig. 8 Traces $p(t, \vec{\mathbf{x}}_r, \vec{\mathbf{x}}_s)$ in the homogeneous medium (*top*) and in one realization of the random medium, for $\sigma = 0.1\,\%$. The traces are similar, except for the wavefront distortion noted in the *bottom plot*

of them are receivers except for the central one, which is also a source. It emits a pulse $f(t)$ that is a modulated sinc function, with constant Fourier coefficients in the bandwidth $[125, 175]$ kHz. The central wavelength is $\lambda_o = 2$ cm.

We display in Fig. 8 the simulated data (traces) $p(t, \vec{\mathbf{x}}_r, \vec{\mathbf{x}}_s)$ at the receiver locations along one cross-range line in the array, with the source in the middle. Each trace is obtained by Fourier synthesis of $\hat{p}(\omega, \vec{\mathbf{x}}_r, \vec{\mathbf{x}}_s)$, for a discretization of the bandwidth in one hundred equidistant frequency intervals of 0.5 kHz. The top plot in Fig. 8 is for the homogeneous medium and the bottom plot for one realization of the random medium, with $\sigma = 0.1\,\%$. The traces look similar, but we note in the bottom plot that the wavefront is distorted.

We show in Fig. 9 the sample estimates of the mean Kirchhoff migration and CINT imaging functions, normalized by the standard deviation at their peak, which is at the reflector location $\vec{\mathbf{y}}$. We display the square root for the CINT image, so we can compare the resolution of the two methods. If we did not have the windows in CINT, the imaging function would be the square of the migration one, which is why we take the square root in Fig. 9. We note that, as predicted by the theory, the CINT image is blurier than the migration one, but it has a much higher SNR at $\vec{\mathbf{y}}$. Kirchhoff migration gives a very small SNR, so the images are expected to change a lot with the realizations of the random medium. The SNR of the CINT image is of order one in this simulation because the aperture a is not very large. Thus, the CINT images change with the realizations as well, but the changes are smaller and never exceed the mean. The peak is expected to stay close to $\vec{\mathbf{y}}$, independent of the realization. This is illustrated in Fig. 10 where we display the migration images (top row) and CINT images (bottom row) in three realizations of the random medium.

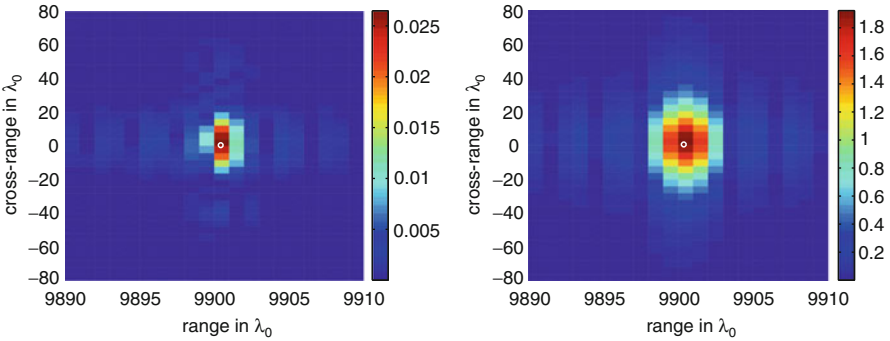


Fig. 9 The mean of the Kirchhoff migration (*left*) and square root of CINT images (*right*) divided by their standard deviation at the peak. The *colorbar* shows that the SNR of CINT is approximately 100 times larger than that of Kirchhoff migration. The axes are range and cross-range in λ_o, measured from the source at the array. The true reflector location is in the middle of the search domain, and it is indicated with the *white circle*

The migration images are quite different from each other: the first image is not as well focused as the others and the peak changes its location. The CINT images change much less with the realization and are very similar to their mean displayed in the right plot in Fig. 9.

9 Appendix 1: Second Moments of the Random Travel Time

We obtain by direct calculation from (75) and (84) that

$$\begin{aligned}
&\mathbb{E}\left[\nu(\vec{\mathbf{x}},\vec{\mathbf{y}})\nu(\vec{\mathbf{x}}',\vec{\mathbf{y}}') = \frac{\mathcal{L}}{\sqrt{2\pi}\ell}\int_0^1 dt\int_0^1 dt'\right]\\
&\mathbb{E}\left[\mu\left(\frac{(1-t)\vec{\mathbf{y}}}{\ell}+\frac{t\vec{\mathbf{x}}}{\ell}\right)\mu\left(\frac{(1-t')\vec{\mathbf{y}}'}{\ell}+\frac{t'\vec{\mathbf{x}}}{\ell}\right)\right]\\
&= \frac{\mathcal{L}}{\sqrt{2\pi}\ell}\int_0^1 dt\int_0^1 dt'\,\mathcal{R}\left[\frac{(t'-t)(\vec{\mathbf{x}}-\vec{\mathbf{y}})}{\ell}+\frac{(1-t')(\vec{\mathbf{y}}'-\vec{\mathbf{y}})+t'(\vec{\mathbf{x}}'-\vec{\mathbf{x}})}{\ell}\right]\\
&= \frac{1}{\sqrt{2\pi}}\int_0^1 dt'\int_{-\mathcal{L}/\ell}^{\mathcal{L}/\ell} d\tilde{t}\,\mathcal{R}\left[\tilde{t}\,\vec{\mathbf{n}}+\frac{(1-t')(\vec{\mathbf{y}}'-\vec{\mathbf{y}})+t'(\vec{\mathbf{x}}'-\vec{\mathbf{x}})}{\ell}\right], \qquad (174)
\end{aligned}$$

where we made the change of variables $(t,t') \rightsquigarrow (\tilde{t},t')$ with

$$t'-t = \frac{\ell}{\mathcal{L}}\tilde{t}.$$

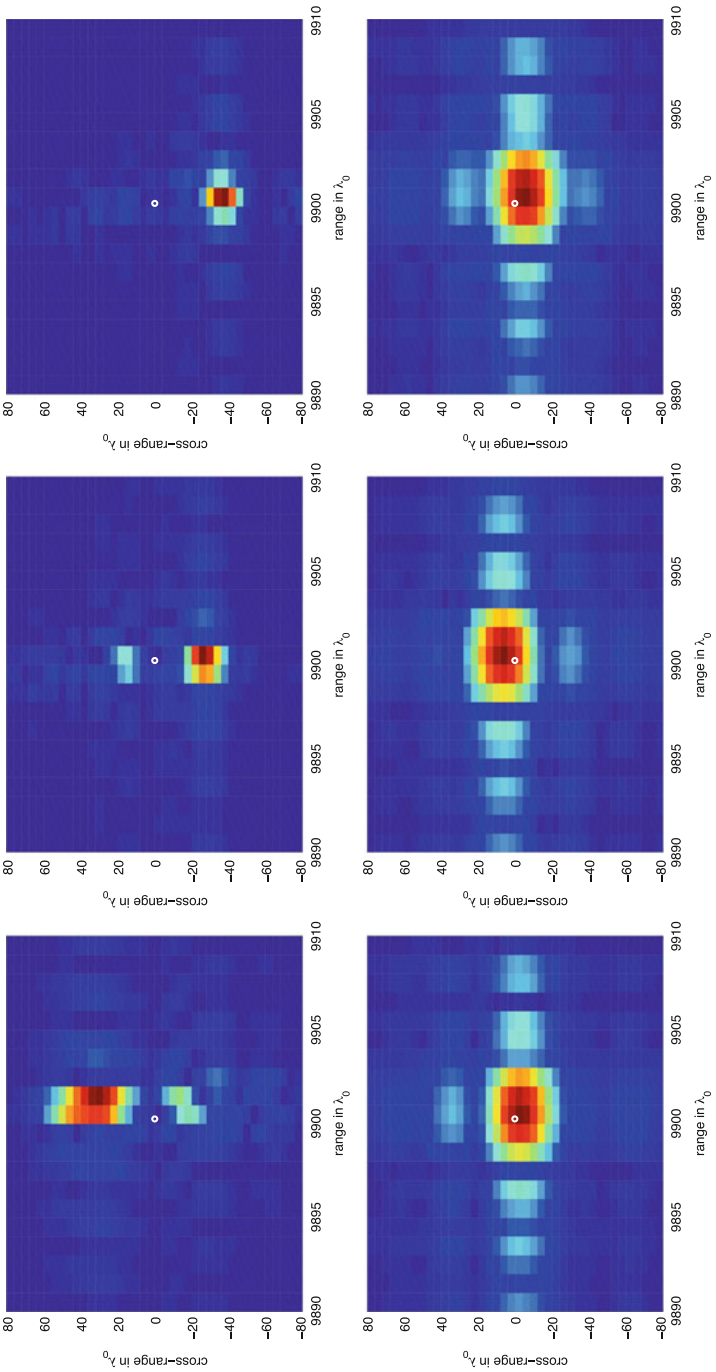


Fig. 10 *Top row*: Kirchhoff migration images in three realizations of the random medium. *Bottom row*: Square root of the CINT images in the same three realizations of the random medium. The axes are range and cross-range in λ_0, measured from the source in the array

We can rewrite the result as

$$\mathbb{E}\left[\nu(\vec{\mathbf{x}},\vec{\mathbf{y}})\nu(\vec{\mathbf{x}}',\vec{\mathbf{y}}')\right] = \int_0^1 dt' \int_{-\infty}^{\infty} \frac{d\tilde{t}}{\sqrt{2\pi}}\, 1_{[-\mathcal{L}/\ell,\mathcal{L}/\ell]}(\tilde{t})\, \mathcal{R}\left[\tilde{t}\,\vec{\mathbf{n}} + \frac{(1-t')(\vec{\mathbf{y}}'-\vec{\mathbf{y}}) + t'(\vec{\mathbf{x}}'-\vec{\mathbf{x}})}{\ell}\right], \tag{175}$$

and note that the integrand converges to $\mathcal{R}\left[\tilde{t}\,\vec{\mathbf{n}} + \frac{(1-t')(\vec{\mathbf{y}}'-\vec{\mathbf{y}})+t'(\vec{\mathbf{x}}'-\vec{\mathbf{x}})}{\ell}\right]$ pointwise, from below, as $\mathcal{L}/\ell \to \infty$. Since $\mathcal{R}$ is integrable by assumption, we obtain from the Lebesgue dominated convergence theorem that

$$\begin{aligned}\mathbb{E}\left[\nu(\vec{\mathbf{x}},\vec{\mathbf{y}})\nu(\vec{\mathbf{x}}',\vec{\mathbf{y}}')\right] &\approx \lim_{\mathcal{L}/\ell\to\infty} \int_0^1 dt' \int_{-\infty}^{\infty} \frac{d\tilde{t}}{\sqrt{2\pi}}\, 1_{[-\mathcal{L}/\ell,\mathcal{L}/\ell]}(\tilde{t}) \\ &\quad \mathcal{R}\left[\tilde{t}\,\vec{\mathbf{n}} + \frac{(1-t')(\vec{\mathbf{y}}'-\vec{\mathbf{y}}) + t'(\vec{\mathbf{x}}'-\vec{\mathbf{x}})}{\ell}\right] \\ &= \int_{-\infty}^{\infty} \frac{d\tilde{t}}{\sqrt{2\pi}}\, \mathcal{R}\left[\tilde{t}\,\vec{\mathbf{n}} + \frac{(1-t')(\vec{\mathbf{y}}'-\vec{\mathbf{y}}) + t'(\vec{\mathbf{x}}'-\vec{\mathbf{x}})}{\ell}\right],\end{aligned} \tag{176}$$

as stated in (86).

10 Appendix 2: Second Moments of the Local Cross-Correlations

Using the Gaussian windows in (147), we obtain

$$\begin{aligned}\mathbb{E}\left[\left|\mathcal{C}(\overline{\tau}_o(\vec{\mathbf{x}},\vec{\mathbf{x}}',\vec{\mathbf{y}}),\tilde{\tau}_o(\vec{\mathbf{x}},\vec{\mathbf{x}}',\vec{\mathbf{y}});\vec{\mathbf{x}},\vec{\mathbf{x}}')\right|^2\right] &= \frac{2\pi}{(32\pi^3L^2)^2B^4}\iint_{-\infty}^{\infty} \\ d\omega_1 d\omega_2\, e^{-\frac{(\omega_1-\omega_o)^2}{B^2}-\frac{(\omega_2-\omega_o)^2}{B^2}} &\quad \times \iint_{-\infty}^{\infty} \\ d\tilde{\omega}_1 d\tilde{\omega}_2 e^{-\frac{(\tilde{\omega}_1^2+\tilde{\omega}_2^2)}{2}\left(\frac{1}{\Omega_\Phi^2}+\frac{1}{2B^2}\right)} & \\ \mathbb{E}\left[e^{i\left(\omega_1-\omega_2+\frac{\tilde{\omega}_1-\tilde{\omega}_2}{2}\right)\nu_\tau(\vec{\mathbf{x}},\vec{\mathbf{y}})-i\left(\omega_1-\omega_2-\frac{\tilde{\omega}_1-\tilde{\omega}_2}{2}\right)\nu_\tau(\vec{\mathbf{x}}',\vec{\mathbf{y}})}\right], &\end{aligned} \tag{177}$$

and it remains to evaluate the expectation. Because ν_τ is approximately Gaussian, we have

$$\mathbb{E}\left[e^{i\left(\Delta\omega+\frac{\Delta\tilde{\omega}}{2}\right)\nu_\tau(\vec{\mathbf{x}},\vec{\mathbf{y}})-i\left(\Delta\omega-\frac{\Delta\tilde{\omega}}{2}\right)\nu_\tau(\vec{\mathbf{x}}',\vec{\mathbf{y}})}\right] \approx e^{-\frac{1}{2}\mathbb{E}\left[\left(\Delta\omega+\frac{\Delta\tilde{\omega}}{2}\right)\nu_\tau(\vec{\mathbf{x}},\vec{\mathbf{y}})-\left(\Delta\omega-\frac{\Delta\tilde{\omega}}{2}\right)\nu_\tau(\vec{\mathbf{x}}',\vec{\mathbf{y}})\right]^2},$$

where we let $\Delta\omega = \omega_1 - \omega_2$ and $\Delta\tilde{\omega} = \tilde{\omega}_1 - \tilde{\omega}_2$. The exponent follows from (84) and (91)

$$\mathbb{E}\left[\Big(\Delta\omega + \frac{\Delta\tilde{\omega}}{2}\Big)\nu_\tau(\vec{\mathbf{x}},\vec{\mathbf{y}}) - \Big(\Delta\omega - \frac{\Delta\tilde{\omega}}{2}\Big)\nu_\tau(\vec{\mathbf{x}}',\vec{\mathbf{y}})\right]^2 \approx \frac{2\big[(\Delta\omega)^2 + \frac{(\Delta\tilde{\omega})^2}{4}\big]}{\Omega^2} \left[1 - \int_0^1 dt\, e^{-\frac{t^2|\mathbf{x}-\mathbf{x}'|^2}{2\ell^2}}\right].$$

Substituting in (177), we obtain

$$\begin{aligned}
&\mathbb{E}\left[\left|\mathcal{C}(\overline{\tau}_o(\vec{\mathbf{x}},\vec{\mathbf{x}}',\vec{\mathbf{y}}),\tilde{\tau}_o(\vec{\mathbf{x}},\vec{\mathbf{x}}',\vec{\mathbf{y}});\vec{\mathbf{x}},\vec{\mathbf{x}}')\right|^2\right] \\
&\quad\approx \frac{2\pi}{(32\pi^3L^2)^2B^4}\iint_{-\infty}^{\infty} d\omega_1 d\omega_2\, e^{-\frac{(\omega_1-\omega_o)^2}{B^2} - \frac{(\omega_2-\omega_o)^2}{B^2}} \\
&\quad\times \iint_{-\infty}^{\infty} d\tilde{\omega}_1 d\tilde{\omega}_2 e^{-\frac{(\tilde{\omega}_1^2+\tilde{\omega}_2^2)}{2}\left(\frac{1}{\Omega_\Phi^2}+\frac{1}{2B^2}\right) - \left[\frac{(\omega_1-\omega_2)^2}{\Omega^2} + \frac{(\tilde{\omega}_1-\tilde{\omega}_2)^2}{4\Omega^2}\right]\left[1-\int_0^1 dt\, e^{-\frac{t^2|\mathbf{x}-\mathbf{x}'|^2}{2\ell^2}}\right]},
\end{aligned} \tag{178}$$

and after evaluating the Gaussian integrals, we obtain (149).

11 Conclusion

This article reviews basic results on coherent array imaging in random media. The random model is motivated by the uncertainty of the small-scale fluctuations of the wave speed in complex media with numerous inhomogeneities. We consider a simple model of wave propagation in random media that captures nevertheless canonical scattering effects on the coherent part of the waves, and consider two imaging methods: migration imaging and CINT imaging. They are both related to an approximation of the solution of the least squares data fit formulation of the inverse problem. Migration imaging is superficially connected to the time reversal process, in the sense that it involves the back-propagation to the imaging region of the time reversed waves measured at the receivers in the array. However, the back-propagation is in a surrogate medium, not in the real one as in the time reversal process, because the medium is not known in imaging. We know only its smooth part, but not its inhomogeneities, which is why we model it as random. This subtle difference between imaging and time reversal has profound effects in random media. We give an explicit and self-contained study of these effects and show that migration imaging is not useful when the waves propagate longer than a scattering mean-free path in random media. The CINT method images by back-propagating to the imaging region local cross-correlations of the measurements at the array. We analyze in detail these local cross-correlations in order to explain why they are useful in

imaging. Moreover, we give an explicit resolution analysis of CINT, which includes an assessment of its statistical stability, and illustrate the results with numerical simulations.

Acknowledgments This article reviews results obtained in collaboration with Josselin Garnier from Université Paris VII, George Papanicolaou from Stanford University, and Chrysoula Tsogka from University of Crete. These results are published in [8, 9, 11, 12]. The work of L. Borcea was partially supported by the AFSOR Grant FA9550-12-1-0117 and the ONR Grant N00014-14-1-0077.

Cross-References

▸ Inverse Scattering
▸ Large-Scale Inverse Problems in Imaging
▸ Synthetic Aperture Radar Imaging
▸ Wave Phenomena

References

1. Alonso, R., Borcea, L., Papanicolaou, G., Tsogka, C.: Detection and imaging in strongly backscattering randomly layered media. Inverse Prob. **27**, 025004 (2011)
2. Aubry, A., Derode, A.: Random matrix theory applied to acoustic backscattering and imaging in complex media. Phys. Rev. Lett. **102**(8), 084301 (2009)
3. Bal, G., Ryzhik, L.: Time reversal and refocusing in random media. SIAM J. Appl. Math. **63**(5), 1475–1498 (2003)
4. Biondi, B.: 3D Seismic Imaging. Society of Exploration Geophysicists, Tulsa (2006)
5. Bleistein, N., Cohen, J.K., Stockwell, J.W.: Mathematics of Multidimensional Seismic Imaging, Migration, and Inversion, vol. 13. Springer, New York (2001)
6. Blomgren, P., Papanicolaou, G., Zhao, H.: Super-resolution in time-reversal acoustics. J. Acoust. Soc. Am. **111**, 230 (2002)
7. Borcea, L., Garnier, J.: Paraxial coupling of propagating modes in three-dimensional waveguides with random boundaries (2012). arXiv:1211.0468. Preprint
8. Borcea, L., Papanicolaou, G., Tsogka, C.: Adaptive interferometric imaging in clutter and optimal illumination. Inverse Prob. **22**, 1405–1436 (2006)
9. Borcea, L., Papanicolaou, G., Tsogka, C.: Asymptotics for the space-time Wigner transform with applications to imaging. In: Baxendale, P.H., Lototsky, S.V., (eds.) Stochastic Differential Equations: Theory and Applications. Volume in Honor of Professor Boris L Rozovskii, Interdisciplinary Mathematical Sciences, vol. 2. World Scientific Publishing Company, Hackensack (2007)
10. Borcea, L., del Cueto, F.G., Papanicolaou, G., Tsogka, C.: Filtering random layering effects in imaging. SIAM Multiscale Model. Simul. **8**, 751–781 (2010)
11. Borcea, L., Garnier, J., Papanicolaou, G., Tsogka, C.: Enhanced statistical stability in coherent interferometric imaging. Inverse Prob. **27**, 085003 (2011)
12. Borcea, L., Garnier, J., Papanicolaou, G., Tsogka, C.: Coherent interferometric imaging, time gating and beamforming. Inverse Prob. **27**(6), 065008 (2011)
13. Borcea, L., Papanicolaou, G., Tsogka, C.: Adaptive time-frequency detection and filtering for imaging in heavy clutter. SIAM J. Imag. Sci. **4**(3), 827–849 (2011)

14. Carazzone, J., Symes, W.: Velocity inversion by differential semblance optimization. Geophysics **56**, 654 (1991)
15. Chai, A., Moscoso, M., Papanicolaou, G.: Robust imaging of localized scatterers using the singular value decomposition and $\ell 1$ minimization. Inverse Prob. **29**(2), 025016 (2013)
16. Curlander, J.C., McDonough, R.N., Synthetic Aperture Radar-Systems and Signal Processing. Wiley, New York (1991)
17. Dussaud, E.A.: Velocity analysis in the presence of uncertainty. Ph.D. thesis, Rice University (2005)
18. Fannjiang, A.C., Solna, K.: Propagation and time reversal of wave beams in atmospheric turbulence. Multiscale Model. Simul. **3**(3), 522–558 (2005)
19. Fink, M.: Time reversed acoustics. Phys. Today **50**, 34 (1997)
20. Fink, M.: Time-reversed acoustics. Sci. Am. **281**(5), 91–97 (1999)
21. Fouque, J.P., Garnier, J., Papanicolaou, G., Sølna, K.: Wave Propagation and Time Reversal in Randomly Layered Media. Springer, New York (2007)
22. Garnier, J., Sølna, K.: Effective fractional acoustic wave equations in one-dimensional random multiscale media. J. Acoust. Soc. Am. **127**, 62 (2010)
23. John, J., Dennis, E., Schnabel, R.B.: Numerical Methods for Unconstrained Optimization and Nonlinear Equations, vol. 16. SIAM, Philadelphia (1983)
24. Lewis, R.M., Keller, J.B.: Asymptotic methods for partial differential equations: the reduced wave equation and Maxwell's equation. Technical report, DTIC Document (1964)
25. Marty, R., Sølna, K.: A general framework for waves in random media with long-range correlations. Ann. Appl. Probab. **21**(1), 115–139 (2011)
26. Moscoso, M., Novikov, A., Papanicolaou G., Ryzhik, L.: A differential equations approach to l1-minimization with applications to array imaging. Inverse Prob. **28**(10), 105001 (2012)
27. Papanicolaou, G., Ryzhik, L., Sølna, K.: Self-averaging from lateral diversity in the itô-schrödinger equation. Multiscale Model. Simul. **6**(2), 468–492 (2007)
28. Rytov, SM., Kravtsov, Y.A., Tatarskii, V.I.: Principles of Statistical Radiophysics. Vol. 4. Wave Propagation Through Random Media. Springer, Berlin (1989)
29. Ryzhik, L., Papanicolaou, G., Keller, J.B.: Transport equations for elastic and other waves in random media. Wave Motion **24**(4), 327–370 (1996)
30. Santosa, F., Symes, W.W., Brown, R.L.: An Analysis of Least-Squares Velocity Inversion. Society of Exploration Geophysicists, Tulsa (1989)
31. Snons C., Sukjoon P., Cs, S.: Two efficient steepest-descent algorithms for source signature-free waveform inversion: synthetic examples. J. Seism. Explor. **14**(4), 335 (2006)
32. Stolk, C.: On the modeling and inversion of seismic data. Ph.D. thesis, Universiteit Utrecht (2000)
33. Symes, WW.: The seismic reflection inverse problem. Inverse Prob. **25**(12), 123008 (2009)
34. Tarantola, A., Crase, E., Jervis, M., Konen, Z., Lindgren, J., Mosegaard, K., Noble, M.: Nonlinear inversion of seismograms: State of the art. In: 60th Annual International Meeting, Expanded Abstracts, pp. 1193–1198. Society of Exploration Geophysicists (1990)
35. van Rossum, M.C.W., Nieuwenhuizen, Th.M.: Multiple scattering of classical waves: microscopy, mesoscopy, and diffusion. Rev. Mod. Phys. **71**(1), 313 (1999)
36. Verschuur, D.J., Berkhout, A.J., Wapenaar, C.P.A.: Adaptive surface-related multiple elimination. Geophysics **57**(9), 1166–1177 (1992)
37. Virieux, J., Operto, S.: An overview of full-waveform inversion in exploration geophysics. Geophysics **74**(6), WCC1–WCC26 (2009)

Part III

Image Restoration and Analysis

Statistical Methods in Imaging

Daniela Calvetti and Erkki Somersalo

Contents

1 Introduction 1344
2 Background 1345
Images in the Statistical Setting 1345
Randomness, Distributions, and Lack of Information 1346
Imaging Problems 1348
3 Mathematical Modeling and Analysis 1350
Prior Information, Noise Models, and Beyond 1350
Accumulation of Information and Priors 1350
Likelihood: Forward Model and Statistical Properties of Noise 1354
Maximum Likelihood and Fisher Information 1358
Informative or Noninformative Priors? 1359
Adding Layers: Hierarchical Models 1359
4 Numerical Methods and Case Examples 1362
Estimators 1362
Algorithms 1368
Statistical Approach: What Is the Gain? 1382
5 Conclusion 1389
Cross-References 1389
References 1390

D. Calvetti (✉)
Department of Mathematics and Department of Cognitive Science, Case Western Reserve University, Cleveland, OH, USA
e-mail: daniela.calvetti@case.edu

E. Somersalo
Department of Mathematics, Case Western Reserve University, Cleveland, OH, USA
e-mail: erkki.somersalo@case.edu

© Springer Science+Business Media New York 2015

O. Scherzer (ed.), *Handbook of Mathematical Methods in Imaging*,
DOI 10.1007/978-1-4939-0790-8_53

Abstract

The theme of this chapter is statistical methods in imaging, with a marked emphasis on the Bayesian perspective. The application of statistical notions and techniques in imaging requires that images and the available data are redefined in terms of random variables, the genesis and interpretation of randomness playing a major role in deciding whether the approach will be along frequentist or Bayesian guidelines. The discussion on image formation from indirect information, which may come from non-imaging modalities, is coupled with an overview of how statistics can be used to overcome the hurdles posed by the inherent ill-posedness of the problem. The statistical counterpart to classical inverse problems and regularization approaches to contain the potentially disastrous effects of ill-posedness is the extraction and implementation of complementary information in imaging algorithms. The difficulty in expressing quantitative and uncertain notions about the imaging problem at hand in qualitative terms, which is a major challenge in a deterministic context, can be more easily overcome once the problem is expressed in probabilistic terms. An outline of how to translate some typical qualitative traits into a format which can be utilized by statistical imaging algorithms is presented. In line with the Bayesian paradigm favored in this chapter, basic principles for the construction of priors and likelihoods are presented, together with a discussion of numerous computational statistics algorithms, including maximum likelihood estimators, maximum a posteriori and conditional mean estimators, expectation maximization, Markov chain Monte Carlo, and hierarchical Bayesian models. Rather than aiming to be a comprehensive survey, the present chapter hopes to convey a wide and opinionated overview of statistical methods in imaging.

1 Introduction

Images, alone or in sequences, provide a very immediate and effective way of transferring information, as the human eye–brain complex is extremely well adapted at extracting quickly their salient features, let them be edges, textures, anomalies, or movement. While the amount of information that can be compressed in an image is tremendously large and varied, the image processing ability of the human eye is so advanced to outperform the most advanced of algorithms. One of the reasons why the popularity of statistical tools in imaging continues to grow is the flexibility that this modality offers when it comes to utilizing qualitative attributes of the images or to recover them from indirect, corrupt specimens. The utilization of qualitative clues to augment scarce data is akin to the process followed by the eye–brain system.

Statistics, which according to Pierre–Simon Laplace, is "common sense expressed in terms of numbers," is well suited for quantifying qualitative attributes. The opportunity to augment poor quality data with complementary information which may be based on our preconception of what we are looking for or on

information coming from sources other than the data makes statistical methods particularly attractive in imaging applications.

In this chapter, we present a brief overview of some of the key concepts and most popular algorithms in statistical imaging, highlighting the similarity and the differences with the closest deterministic counterparts. A particular effort is made to demonstrate that the statistical methods lead to new ideas and algorithms that the deterministic methods do not give.

2 Background

Images in the Statistical Setting

The mathematical vessel that we will use here to describe a black and white image is a matrix with nonnegative entries, each representing the light intensity at one pixel of the discretized image. Color images can be thought of as the result of superimposing a few color intensity matrices; in most application, a color image is represented by three matrices, for example, encoding the red, green, and blue intensity at each pixel. While color imaging applications can also be approached with statistical methods, here we will only consider gray-scale images. Thus, an image X is represented as a matrix

$$\mathsf{X} = \left[x_{ij}\right], \quad 1 \le i \le n,\ 1 \le j \le m,\ x_{ij} \ge 0.$$

In our treatment, we will not worry about the range of the image pixel values, assuming that, if necessary, the values are appropriately normalized. Notice that this representation tacitly assumes that we restrict our discussion to rectangular images discretized into rectangular arrays of pixels. This hypothesis is neither necessary nor fully justified, but it simplifies the notation in the remainder of the chapter. In most imaging algorithms, the first step consists of storing the image into a vector by reshaping the rectangular matrix. We use here a columnwise stacking, writing

$$\mathsf{X} = \left[x^{(1)}\ x^{(2)}\ \dots\ x^{(m)}\right], \quad x^{(j)} \in \mathbb{R}^n,\ 1 \le j \le m,$$

and further

$$x = \mathrm{vec}\,(\mathsf{X}) = \begin{bmatrix} x^{(1)} \\ \vdots \\ x^{(m)} \end{bmatrix} \in \mathbb{R}^N, \quad N = n \times m.$$

Images can be either directly observed or represent a function of interest, as is, for example, the case for tomographic images.

Randomness, Distributions, and Lack of Information

We start this section by introducing some notations. A multivariate random variable $X : \Omega \to \mathbb{R}^N$ is a measurable mapping from a probability space Ω equipped with a σ-algebra and a probability measure P. The elements of $\mathbb{R}^N$, as well as the realizations of X, are denoted by lowercase letters, that is, for $\omega \in \Omega$ given, $X(\omega) = x \in \mathbb{R}^N$. The probability distribution μ_X is the measure defined as

$$\mu_X(B) = \mathsf{P}\left(X^{-1}(B)\right), \quad B \subset \mathbb{R}^N \text{ measurable.}$$

If μ_X is absolutely continuous with respect to the Lebesgue measure, there is a measurable function π_X, the Radon–Nikodym derivative of μ_X with respect to the Lebesgue measure such that

$$\mu_X(B) = \int_B \pi_X(x)dx.$$

For the sake of simplicity, we shall assume that all the random variables define probability distributions which are absolutely continuous with respect to the Lebesgue measure.

Consider two random variables $X : \Omega \to \mathbb{R}^N$ and $Y : \Omega \to \mathbb{R}^M$. The joint probability density is defined first over Cartesian products,

$$\mu_{X,Y}(B \times D) = \mathsf{P}\left(X^{-1}(B) \cap Y^{-1}(D)\right),$$

and then extended to the whole product σ-algebra over $\mathbb{R}^N \times \mathbb{R}^M$. Under the assumption of absolute continuity, the joint density can be written as

$$\mu_{X,Y}(B \times D) = \int_B \int_D \pi_{X,Y}(x, y)dydx,$$

where $\pi_{X,Y}$ is a measurable function. This definition extends naturally to the case of more than two random variables.

Since the notation just introduced here gets quickly rather cumbersome, we will simplify it by dropping the subscripts, writing $\pi_{X,Y}(x, y) = \pi(x, y)$, that is, letting x and y be at the same time variables and indicators of their parent uppercase random variables. Furthermore, since the ordering of the random variables is irrelevant – indeed, $\mathsf{P}\left(X^{-1}(B) \cap Y^{-1}(D)\right) = \mathsf{P}\left(Y^{-1}(D) \cap X^{-1}(B)\right)$ – we will occasionally interchange the roles of x and y in the densities, without assuming that the probability densities should be symmetric in x and y. In other words, we will use π as a generic symbol for "probability density."

With these notations, given two random variables X and Y, define the marginal densities

$$\pi(x) = \int_{\mathbb{R}^M} \pi(x, y)dy, \quad \pi(y) = \int_{\mathbb{R}^N} \pi(x, y)dx,$$

which express the probability densities of X and Y, respectively, on their own, while the other variable is allowed to take on any value. By fixing y, and assuming that $\pi(y) \neq 0$, we have that

$$\int_{\mathbb{R}^N} \frac{\pi(x, y)}{\pi(y)} dx = 1;$$

hence, the nonnegative function

$$x \mapsto \pi(x \mid y) \stackrel{\text{def}}{=} \frac{\pi(x, y)}{\pi(y)} \tag{1}$$

defines a probability distribution for X referred to as the conditional density of X, given $Y = y$. Similarly, we define the conditional density of Y given $X = x$ as

$$\pi(y \mid x) \stackrel{\text{def}}{=} \frac{\pi(x, y)}{\pi(x)}. \tag{2}$$

This rather expedite way of defining the conditional densities does not fully explain why this interpretation is legitimate; a more rigorous explanation can be found in textbooks on probability theory [8, 18].

The concept of probability measure does not require any further interpretation to yield a meaningful framework for analysis, and this indeed is the viewpoint of theoretical probability. When applied to real-world problems, however, an interpretation is necessary, and this is exactly where the opinions of statisticians start to diverge. In frequentist statistics, the probability of an event is its asymptotic relative frequency of occurrence as the number of repeated experiments tend to infinity, and the probability density can be thought of as a limit of histograms. A different interpretation is based on the concept of information. If the value of a quantity is either known or it is at potentially retrievable from the available information, there is no need to leave the deterministic realm. If, on the other hand, the value of a quantity is uncertain in the sense that the available information is insufficient to determine it, to view it as a random variable appears natural. In this interpretation of randomness, it is immaterial whether the lack of information is contingent ("imperfect measurement device, insufficient sampling of data") or fundamental ("quantum physical description of an observable"). It should also be noted that the information, and therefore the concept of probability, is subjective, as the value of a quantity may be known to one observer and unknown to another [14, 18]. Only in the latter case the concept of probability is needed. The interpretation of probability in this chapter follows mostly the subjective, or Bayesian tradition, although most of the time the distinction is immaterial. Connections to non-Bayesian statistics are made along the discussion.

Most imaging problems can be recast in the form of a statistical inference problem. Classically, inverse problems are stated as follows: *Given an observation of a vector $y \in \mathbb{R}^M$, find an estimate of the vector $x \in \mathbb{R}^N$, based on the forward*

model mapping x to y. Statistical inference, on the other hand, is concerned with identifying a probability distribution that the observed data is presumably drawn from. In the frequentist statistics, the observation y is seen as a realization of a random variable Y, the unknown x being a deterministic parameter that determines the underlying distribution $\pi(y \mid x)$, or *likelihood density*, and hence the estimation of x is the object of interest. In contrast, in the Bayesian setting, both variables x and y are first extended to random variables, Y and X, respectively, as discussed in more detail in the following sections. The marginal density $\pi(x)$, which is independent of the observation y, is called the *prior density* and denoted by $\pi_{\text{prior}}(x)$, while the likelihood is the conditional density $\pi(y \mid x)$. Combining the formulas (1) and (2), we obtain

$$\pi(x \mid y) = \frac{\pi_{\text{prior}}(x)\pi(y \mid x)}{\pi(y)},$$

which is the celebrated Bayes' formula [3]. The conditional distribution $\pi(x \mid y)$ is the *posterior distribution* and, in the Bayesian statistical framework, the solution of the inverse problem.

Imaging Problems

A substantial body of classical imaging literature is devoted to problems where the data consists of an image, represented here as a vector $y \in \mathbb{R}^M$ that is either a noisy, blurred, or otherwise corrupt version of the image $x \in \mathbb{R}^N$ of primary interest. The canonical model for this class of imaging problems is

$$y = \mathsf{A}x + \text{“noise,”} \tag{3}$$

where the properties of the matrix A depend on the imaging problem at hand. A more general imaging problems is of the form

$$y = F(x) + \text{“noise,”} \tag{4}$$

where the function $F : \mathbb{R}^N \mapsto \mathbb{R}^M$ may be a nonlinear function and the data y need not even represent an image. This is a common setup in medical imaging applications with a nonlinear forward model.

In classical, nonstatistical framework, imaging problems, and more generally, inverse problems, are often, somewhat arbitrarily, classified as being linear or nonlinear, depending on whether the forward model F in (4) is linear or nonlinear. In the statistical framework, this classification is rather irrelevant. Since probability densities depend not only on the forward map but also on the noise and, in the Bayesian case, the prior models, even a linear forward map can result in a nonlinear

estimation problem. We review some widely studied imaging problems to highlight this point.

1. *Denoising*: Denoising refers to the problem of removing noise from an image which is otherwise deemed to be a satisfactory representation of the information. The model for denoising can be identified with (3), with $M = N$ and the identity $\mathsf{A} = \mathsf{I} \in \mathbb{R}^{N\times N}$ as forward map.
2. *Deblurring*: Deblurring is the process of removing a blur, due, for example, to an imaging device being out of focus, to motion of the object during imaging ("motion blur"), or to optical disturbances in atmosphere during image formation. Since blurred images are often contaminated by exogenous noise, denoising is an integral part of the deblurring process. Given the image matrix $\mathsf{X} = [x_{ij}]$, the blurring is usually represented as

$$y_{ij} = \sum_{k,\ell} a_{ij,k\ell} x_{k\ell} + \text{"noise."}$$

   Often, but not without loss of generality, the blurring matrix can be assumed to be a convolution kernel,

$$a_{ij,k\ell} = a_{i-k,j-\ell},$$

   with the obvious abuse of notations. It is a straightforward matter to arrange the elements, so that the above problem takes on the familiar matrix–vector form $y = \mathsf{A}x$, and in the presence of noise, the model coincides with (3).
3. *Inpainting*: Here, it is assumed that part of the image x is missing due to an occlusion, a scratch, or other damages. The problem is to paint in the occlusion based on the visible part of the image. In this case, the matrix A in the linear model (3) is a sampling matrix, picking only those pixels of $x \in \mathbb{R}^N$ that are present in $y \in \mathbb{R}^M$, $M < N$.
4. *Image formation*: Image formation is the process of translating data into the form of an image. The process is common in medical imaging, and the description of the forward model connecting the sought image to data may involve linear or nonlinear transformations. An example of a linear model arises in tomography: The image is explored one line at the time, in the sense that the data consist of line integrals indirectly measuring the amount of radiation absorbed in the trajectory from source to detector or the number of photons emitted at locations along the trajectory between pairs of detectors. The problem is of the form (3). An example of a nonlinear imaging model (4) arises in near-infrared optical tomography, in which the object of interest is illuminated by near-infrared light sources, and the transmitted and scattered light intensity is measured in order to form an image of the interior optical properties of the body.

Some of these examples will be worked out in more details below.

3 Mathematical Modeling and Analysis

Prior Information, Noise Models, and Beyond

The goal in Bayesian statistical methods in imaging is to identify and explore probability distributions of images rather than looking for single images, while in the non-Bayesian framework, one seeks to infer on deterministic parameter vectors defining the distribution that the observations are drawn from. The main player in non-Bayesian statistics is the likelihood function, in the notation of section "Randomness, Distributions and Lack of Information," $\pi(y \mid x)$, where $y = y_{\text{observed}}$. In Bayesian statistics, the focus is on the posterior density $\pi(x \mid y)$, $y = y_{\text{observed}}$, the likelihood function being a part of it as indicated by Bayes' formula.

We start the discussion with the Bayesian concept of prior distribution, the non-Bayesian modeling paradigm being discussed in connection with the likelihood function.

Accumulation of Information and Priors

To the question, what should be in a prior for an imaging problem, the best answer is whatever can be built using available information about the image which can supplement the measured data. The information to be accounted by the prior can be gathered in many different ways. Any visually relevant characteristic of the sought image is suitable for a prior, including but not limited to texture, light intensity, and boundary structure. Although it is often emphasized that in a strict Bayesian framework the prior and the likelihood must be constructed separately, in several imaging problems, the setup may be impractical, and the prior and likelihood need to be set up simultaneously. This is the case, for example, when the noise is correlated with the signal itself. Furthermore, some algorithms may contain intermediate steps that formally amount to updating of the a priori belief, a procedure that may seem dubious in the traditional formal Bayesian setting but can be justified in the framework of hierarchical models. For example, in the restoration of images with sharp contrasts from severely blurred, noisy copies, an initially very vague location of the gray-scale discontinuities can be made more precise by extrapolation from intermediate restorations, leading to a Bayesian learning model.

It is important to understand that in imaging, the use of complementary information to improve the performance of the algorithms at hand is a very natural and widespread practice and often necessary to link the solution of the underlying mathematical problem to the actual imaging application. There are several constituents of an image that are routinely handled under the guidance of a priori belief even in fully deterministic settings. A classical example is the assignment of

boundary conditions for an image, a problem which has received a lot of attention over the span of a couple of decades (see, e.g., [21] and references therein). In fact, since it is certainly difficult to select the most appropriate boundary condition for a blurred image, ultimately the choice is based on a combination of a priori belief and algorithmic considerations. The implementation of boundary conditions in deterministic algorithms can therefore be interpreted as using a prior, expressing an absolute belief in the selected boundary behavior. The added flexibility which characterizes statistical imaging methodologies makes it possible to import in the algorithm the postulated behavior of the image at the boundary with a certain degree of uncertainty.

The distribution of gray levels within an image and the transition between areas with different gray-scale intensities are the most likely topics of a priori beliefs, hence primary targets for priors. In the nonstatistical imaging framework, a common choice of regularization, for the underlying least squares problems is a regularization functional, which penalizes growth in the norm of the derivative of the solution, thus discouraging solutions with highly oscillatory components. The corresponding statistical counterpart is a Markov model, based, for example, on the prior assumption that the gray-scale intensity at each pixel is a properly weighted average of the intensities of its neighbors plus a random innovation term which follows a certain statistical distribution. As an example, assuming a regular quadrilateral grid discretization, the typical local model can be expressed in terms of probability densities of pixel values X_j conditioned on the values of its neighboring pixels labeled according to their relative position to X_j as X_{up}, X_{down}, X_{left}, and X_{right}, respectively. The conditional distribution is derived by writing

$$\begin{aligned} X_j \,\big|\, (X_{\text{up}} = x_{\text{up}}, X_{\text{down}} = x_{\text{down}}, X_{\text{left}} = x_{\text{left}}, X_{\text{right}} = x_{\text{right}}) \\ = \frac{1}{4}(x_{\text{up}} + x_{\text{down}} + x_{\text{left}} + x_{\text{right}}) + \Phi_j, \end{aligned} \tag{5}$$

where Φ_j is a random innovation process. For boundary pixels, an appropriate modification reflecting the a priori belief of the extension of the image outside the field of view must be incorporated. In a large variety of application, Φ_j is assumed to follow a normal distribution

$$\Phi_j \sim \mathcal{N}\left(0, \sigma_j^2\right),$$

the variance σ_j^2 reflecting the expected deviation from the average intensity of the neighboring pixels. The Markov model can be expressed in matrix–vector form as

$$\mathsf{L}X = \Phi,$$

where the matrix L is the five-point stencil discretization of the Laplacian in two dimensions and the vector $\Phi \in \mathbb{R}^N$ contains the innovation terms Φ_j. As we assume the innovation terms to be independent, the probability distribution of Φ is

$$\Phi \sim \mathcal{N}(0, \Sigma), \quad \Sigma = \begin{bmatrix} \sigma_1^2 & & & \\ & \sigma_2^2 & & \\ & & \ddots & \\ & & & \sigma_N^2 \end{bmatrix},$$

and the resulting prior model is a *second-order Gaussian smoothness prior*,

$$\pi_{\text{prior}}(x) \propto \exp\left(-\frac{1}{2}\|\Sigma^{-1/2}\mathsf{L}x\|^2\right).$$

Observe that the variances σ_j^2 allow a spatially inhomogeneous a priori control of the texture of the image. Replacing the averaging weights $1/4$ in (5) by more general weights p_k, $1 \le k \le 4$ leads to a smoothness prior with directional sensitivity. Random draws from such anisotropic Gaussian priors are shown in Fig. 1, where each pixel with coordinate vector r_j in a quadrilateral grid has eight neighboring pixels with coordinates r_j^k, and the corresponding weights p_k are chosen as

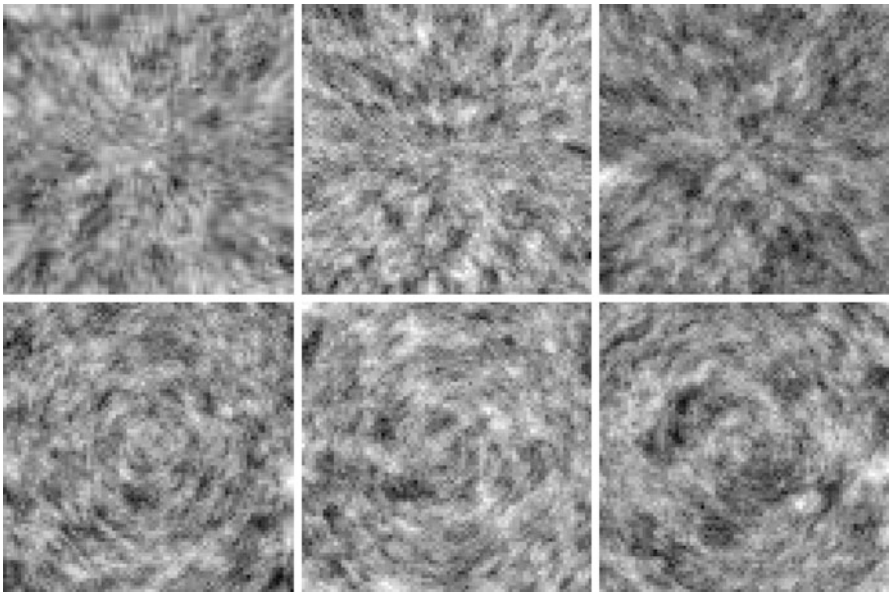


Fig. 1 Random draws from anisotropic Markov models. In the *top row*, the Markov model assumes stronger dependency between neighboring pixels in the radial than in angular direction, while in the *bottom row*, the roles of the directions are reversed. See text for a more detailed discussion

$$p_k = \frac{1}{\tau} \frac{\left(v_j^{\mathsf{T}} \left(r_j - r_j^k \right) \right)^2}{\left| r_j - r_j^k \right|^2}, \quad \tau = 1.1,$$

and the unit vector v_j is chosen either as a vector pointing out of the center of the image (top row) or in a perpendicular direction (bottom row). The former choice thus assumes that pixels are more strongly affected by the adjacent values in the radial direction, while in the latter case, they have less influence than those in the angular direction. The factor τ is added to make the matrix diagonally dominated.

The just described construction of the smoothness prior is a particular instance of priors based on the assumption that the image is a *Markov random field*, (MRF). Similarly to the four-point average example, Markov random fields assume that the conditional probability distribution of a single pixel value X_j conditioned on the remaining image depends only on the neighbors of X_j,

$$\pi \left(x_j \mid x_k, k \neq j \right) = \pi \left(x_j \mid x_k \in N_j \right),$$

where N_j is the list of neighbor pixels of X_j, such as the four adjacent pixels in the model (5). In fact, the *Hammersley–Clifford theorem* (see [5]) states that prior distributions of MRF models are of the form

$$\pi_{\text{prior}}(x) \propto \exp\left(-\sum_{j=1}^{N} V_j(x) \right),$$

where the function $V_j(x)$ depends only on x_j and its neighbors. The simplest model in this family is a Gaussian white noise prior, where $N_j = \emptyset$ and $V_j(x) = x_j^2/(2\sigma^2)$, that is,

$$\pi_{\text{prior}}(x) \propto \exp\left(-\frac{1}{2\sigma^2} \|x\|^2 \right).$$

Observe that this prior assumes mutual independency of the pixels, which has qualitative repercussions on the images based on it.

There is no theoretical reason to restrict the MRFs to Gaussian fields, and in fact, some of the non-Gaussian fields have had a remarkable popularity and success in the imaging context. Two non-Gaussian priors are particularly worth mentioning here, the ℓ^1-prior, where $N_j = \emptyset$ and $V_j(x) = \alpha |x_j|$, that is,

$$\pi_{\text{prior}}(x) \propto \exp\left(-\alpha \|x\|_1 \right), \quad \|x\|_1 = \sum_{j=1}^{N} |x_j|,$$

and the closely related total variation (TV) prior,

$$\pi_{\text{prior}}(x) \propto \exp\left(-\alpha \text{TV}(x)\right), \quad \text{TV}(x) = \sum_{j=1}^{N} V_j(x),$$

with

$$V_j(x) = \frac{1}{2} \sum_{k \in N_j} |x_j - x_k|.$$

The former is suitable for imaging sparse images, where all but few pixels are believed to coincide with the background level that is set to zero. The latter prior is particularly suitable for blocky images, that is, for images consisting of piecewise smooth simple shapes. There is a strong connection to the recently popular concept of *compressed sensing*, see, for example, [11].

MRF priors, or priors with only local interaction between pixels, are by far the most commonly used priors in imaging. It is widely accepted and to some extent demonstrated (see [6] and the discussion in it) that the posterior density is sensitive to local properties of the prior only, while the global properties are predominantly determined by the likelihood. Thus, as far as the role of priors is concerned, it is important to remember that until the likelihood is taken into account, there is no connection with the measured data, hence no reason to believe that the prior should generate images that in the large scale resemble what we are looking for. In general, priors are usually designed to carry very general often qualitative and local information, which will be put into proper context with the guidance of the data through the integration with the likelihood. To demonstrate the local structure implied by different priors, in Fig. 2, we show some random draws from the priors discussed above.

Likelihood: Forward Model and Statistical Properties of Noise

If an image is worth a thousand words, a proper model of the noise corrupting it is worth at least a thousand more, in particular when the processing is based on the statistical methods. So far, the notion of noise has remained vague, and its role unclear. It is the noise, and in fact its statistical properties, that determines the likelihood density. We start by considering two very popular noise models.

Additive, nondiscrete noise: An additive noise model assumes that the data and the unknown are in a functional relation of the form

$$y = F(x) + e, \tag{6}$$

where e is the noise vector. If the function F is linear, or it has been linearized, the problem simplifies to

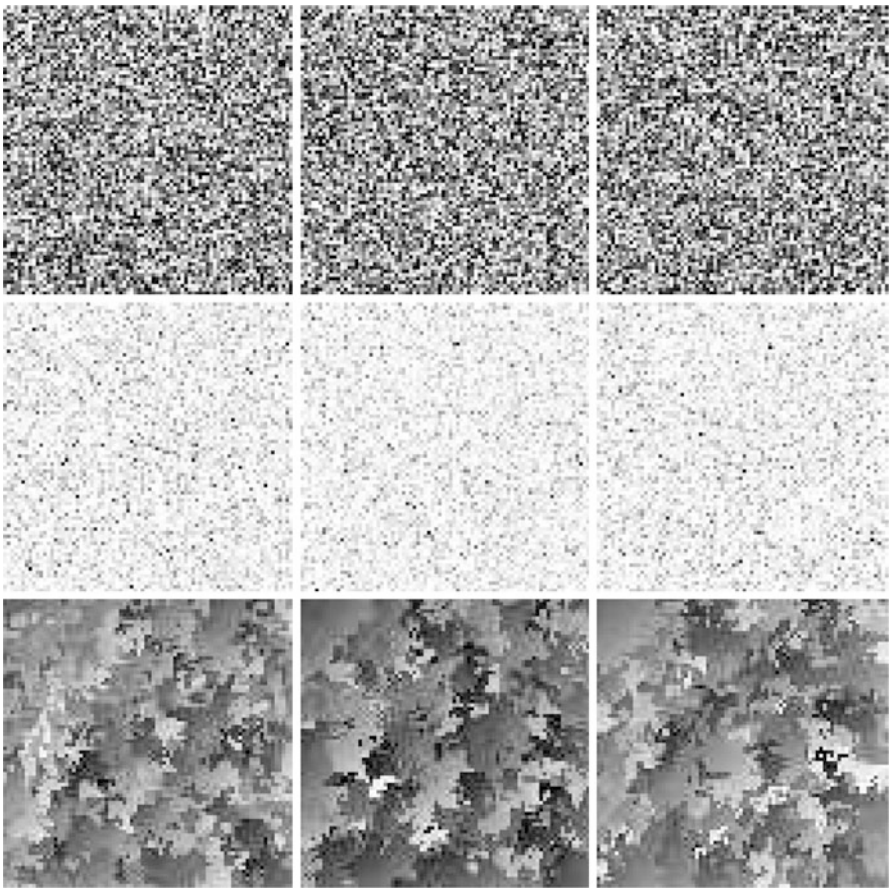


Fig. 2 Random draws from various MRF priors. *Top row*: white noise prior. *Middle row*: sparsity prior or ℓ^1-prior with positivity constraint. *Bottom row*: total variation prior

$$y = \mathsf{A}x + e. \tag{7}$$

The stochastic extension of (6) is

$$Y = F(X) + E,$$

where Y, X, and E are multivariate random vectors.

The form of the likelihood is determined not only by the assumed probability distributions of Y, X, and E but also by the dependency between pairs of these variables. In the simplest case, X and E are assumed to be mutually independent and the probability density of the noise vector known,

$$E \sim \pi_{\text{noise}}(e),$$

resulting in a likelihood function of the form

$$\pi(y \mid x) \propto \pi_{\text{noise}}\,(y - F(x))\,,$$

which is one of the most commonly used in applications. A particularly popular model for additive noise is a Gaussian noise,

$$E \sim \mathcal{N}(0, \Sigma),$$

where the covariance matrix Σ is positive definite. Therefore, if we write $\Sigma^{-1} = \mathsf{D}^{\mathsf{T}}\mathsf{D}$, where D can be the Cholesky factor of Σ^{-1} or $\mathsf{D} = \Sigma^{-1/2}$, the likelihood can be written as

$$\begin{aligned}\pi(y \mid x) &\propto \exp\left(-\frac{1}{2}(y - F(x))^{\mathsf{T}}\Sigma^{-1}(y - F(x))\right)\\ &= \exp\left(-\frac{1}{2}\|\mathsf{D}(y - F(x))\|^2\right).\end{aligned} \tag{8}$$

In the general case where X and E are not independent, we need to specify the joint density

$$(X, E) \sim \pi(x, e)$$

and the corresponding conditional density

$$\pi_{\text{noise}}(e \mid x) = \frac{\pi(x, e)}{\pi_{\text{prior}}(x)}.$$

In this case, the likelihood becomes

$$\pi(y \mid x) \propto \pi_{\text{noise}}(y - F(x) \mid x).$$

This clearly demonstrates the problems which may arise if we want to adhere to the claim that "likelihood should be independent of the prior." Because the interdependency of the image x and the noise is much more common than we might be inclined to believe, the independency of noise and signal is often in conflict with reality. An instance of such situation occurs in electromagnetic brain imaging using magnetoencephalography (MEG) or electroencephalography (EEG), when the eye muscle during a visual task acts as noise source but can hardly be considered as independent from the brain activation due to a visual stimulus. Another example related to boundary conditions will be discussed later on. Also, since the noise term should account not only for the exogenous measurement noise but also for the shortcomings of the model, including discretization errors, the interdependency is in fact a ubiquitous phenomenon too often neglected.

Most additive noise models assume that the noise follows a Gaussian distribution, with zero mean and given covariance. The computational advantages of a Gaussian likelihood are rather formidable and have been a great incentive to use Gaussian approximations of non-Gaussian densities. While it is commonplace and somewhat justified, for example, to approximate Poisson densities with Gaussian densities when the mean is sufficiently large [14], there are some important imaging applications where the statistical distribution of the noise must be faithfully represented in the likelihood.

Counting noise: The weakness of a signal can complicate the deblurring and denoising problem, as is the case in some image processing applications in astronomy [49, 57, 63], microscopy [45, 68], and medical imaging [29, 60]. In fact, in the case of weak signals, a charge-coupled device (CCD), instead of recording an integrated signal over a time window, counts individual photons or electrons. This leads to a situation where the noise corrupting the recorded signal is no longer exogenous but rather an intrinsic property of the signal itself, that is, the input signal itself is a random process with an unpredictable behavior. Under rather mild assumptions – stationarity, independency of increments, and zero probability of coincidence – it can be shown (see, e.g., [62]) that the counting signal follows a Poisson distribution. Consider, for example, the astronomical image of a very distant object, collected with an optical measurement device whose blurring is described by a matrix A. The classical description of such data would follow (7), with the error term collecting the background noise and the thermal noise of the device. The corresponding counting model is

$$y_j \sim \text{Poisson}\left((\mathsf{A}x)_j + b\right),\ y_j, y_k \text{ independent if } j \neq k,$$

or, explicitly,

$$\pi(y \mid x) = \prod_{j=1}^{m} \frac{((\mathsf{A}x)_j + b)^{y_j}}{(y_j)!} \exp\left(-(\mathsf{A}x)_j + b\right),$$

where $b \geq 0$ is a background radiation level, assumed known. Observe that while the data are counts, therefore integer numbers, the expectation need not to be.

Similar or slightly modified likelihoods can be used to model the positron emission tomography (PET) and single-photon emission computed tomography (SPECT) signals; see [29, 54].

The latter example above demonstrates clearly that the description of imaging problems as linear or nonlinear, without a specification of the noise model, in the context of statistical methods, does not play a significant role: Even if the expectation is linear, traditional algorithms for solving linear inverse problems are useless, although they may turn out to be useful within iterative solvers for solving locally linearized steps.

Maximum Likelihood and Fisher Information

When switching to a parametric non-Bayesian framework, the statistical inference problem amounts to estimating a deterministic parameter that identifies the probability distribution from which the observations are drawn. To apply this framework in imaging problems, the underlying image x, which in the Bayesian context was itself a random variable, can be thought of as a parameter vector that specifies the likelihood function,

$$f(y;x) = \pi(y \mid x),$$

as implied by the notation $f(y;x)$ also.

In the non-Bayesian interpretation, a measure of how much information about the parameter x is contained in the observation is given in terms of the *Fisher information matrix* J,

$$\mathsf{J}_{j,k} = \mathsf{E}\left\{\frac{\partial \log f}{\partial x_j}\frac{\partial \log f}{\partial x_k}\right\} = \int \frac{\partial \log f(y;x)}{\partial x_j}\frac{\partial \log f(y;x)}{\partial x_k} f(y;x)dy. \tag{9}$$

In this context, the observation y only is a realization of a random variable Y, whose probability distribution is entirely determined by the distribution of the noise. The gradient of the logarithm of the likelihood function is referred to as the *score*, and the Fisher information matrix is therefore the covariance of the score.

Assuming that the likelihood is twice continuously differentiable and regular enough to allow the exchange of integration and differentiation, it is possible to derive another useful expression for the information matrix. It follows from the identity

$$\frac{\partial \log f}{\partial x_k} = \frac{1}{f}\frac{\partial f}{\partial x_k}, \tag{10}$$

that we may write the Fisher information matrix as

$$\mathsf{J}_{j,k} = \int \frac{\partial \log f}{\partial x_j}\frac{\partial f}{\partial x_k}dy = \frac{\partial}{\partial x_k}\int \frac{\partial \log f}{\partial x_j} f dy - \int \frac{\partial^2 \log f}{\partial x_j \partial x_k} f dy.$$

Using the identity (10) with k replaced by j, we observe that

$$\int \frac{\partial \log f}{\partial x_j} f dy = \int \frac{\partial f}{\partial x_j} dy = \frac{\partial}{\partial x_j}\int f dy = 0,$$

since the integral of f is one, which leads us to the alternative formula

$$\mathsf{J}_{j,k} = -\int \frac{\partial^2 \log f}{\partial x_j \partial x_k} f dy = -\mathsf{E}\left\{\frac{\partial^2 \log f}{\partial x_j \partial x_k}\right\}. \tag{11}$$

The Fisher information matrix is closely related to non-Bayesian estimation theory. This will be discussed later in connection with maximum likelihood estimation.

Informative or Noninformative Priors?

Not seldom the use of priors in imaging applications is blamed for biasing the solution in a direction not supported by the data. The concern of the use of committal priors has led to the search of "noninformative priors" [39] or weak priors that would "let the data speak."

The strength or weakness of a prior is a rather elusive concept, as the importance of the prior in Bayesian imaging is in fact determined by the likelihood: the more information we have about the image in data, the less has to be supplied by the prior. On the other hand, in imaging applications where the likelihood is built on very few data points, the prior needs to supply the missing information, hence has a much more important role. As pointed out before, it is a common understanding that in imaging applications, prior should carry small-scale information about the image that is missing from the likelihood that in turn carries information about the large-scale features and in that sense complements the data.

Adding Layers: Hierarchical Models

Consider the following simple denoising problem with additive Gaussian noise,

$$Y = X + N, \quad N \sim \mathcal{N}(0, \Sigma),$$

with noise covariance matrix Σ presumed known, whose likelihood model is tantamount to saying that

$$Y \mid X = x \sim \mathcal{N}(x, \Sigma).$$

From this perspective, the denoising problem is reduced to estimating the mean of a Gaussian density in the non-Bayesian spirit, and the prior distribution is a *hierarchical* model, expressing the degree of uncertainty of the mean x.

Parametric models are common when defining the prior densities, but similarly to the above interpretation of the likelihood, the parameters are often poorly known. For example, when introducing a prior

$$X \sim \mathcal{N}(\theta, \Gamma)$$

with unknown θ, we are expressing a qualitative prior belief that "X differs from an unknown value by an error with a given Gaussian statistics," which says very little about the values of X itself unless information about θ is provided. Similarly as in the denoising problem, it is natural to augment the prior with another layer of

information concerning the parameter θ. This layering of the inherent uncertainty is at the core of *hypermodels*, or Bayesian hierarchical models. Hierarchical models are not restricted to uncertainties in the prior, but can be applied to lack of information of the likelihood model as well.

In hierarchical models, both the likelihood and the prior may depend on additional parameters,

$$\pi(y \mid x) \to \pi(y \mid x, \gamma), \quad \pi_{\text{prior}}(x) \to \pi_{\text{prior}}(x \mid \theta),$$

with both parameters γ and θ poorly known. In this case, it is natural to augment the model with *hyperpriors*. Assuming for simplicity that the parameters γ and θ are mutually independent so that we can define the hyperprior distributions $\pi_1(\gamma)$ and $\pi_2(\theta)$, the joint probability distribution of all the unknowns is

$$\pi(x, y, \theta, \gamma) = \pi(y \mid x, \gamma)\pi_{\text{prior}}(x \mid \theta)\pi_1(\gamma)\pi_2(\theta).$$

From this point on, the Bayesian inference can proceed along different paths. It is possible to treat the hyperparameters as nuisance parameters and marginalize them out by computing

$$\pi(x, y) = \int\int \pi(x, y, \theta, \gamma) d\theta d\gamma$$

and then proceed as in a standard Bayesian inference problem. Alternatively, the hyperparameters can be included in the list of unknowns of the problem and their posterior density

$$\pi(\xi \mid y) = \frac{\pi(x, y, \theta, \gamma)}{\pi(y)}, \quad \xi = \begin{bmatrix} x \\ \theta \\ \gamma \end{bmatrix}$$

needs to be explored. The estimation of the hyperparameters can be based on the optimization or on the *evidence*, as will be illustrated below with a specific example.

To clarify the concept of a hierarchical model itself, we consider some examples where hierarchical models arise naturally.

Blind deconvolution: Consider the standard deblurring problem defined in section "Imaging Problems." Usually, it is assumed that the blurring kernel A is known, and the likelihood, with additive Gaussian noise with covariance Σ, becomes

$$\pi(y \mid x) \propto \exp\left(-\frac{1}{2}(y - \mathsf{A}x)^{\mathsf{T}}\Sigma^{-1}(y - \mathsf{A}x)\right). \tag{12}$$

In some cases, although A is poorly known, its parametric expression is known and the uncertainty only affects the values of some parameters, as is the case when the shape of the continuous convolution kernel $a(r - s)$ is known but the actual width is not. If we express the kernel a as a function of a width parameter,

$$a(r-s) = a_\gamma(r-s) = \frac{1}{\gamma} a_1(\gamma(r-s)), \quad \gamma > 0,$$

and denote by A_γ the corresponding discretized convolution matrix, the likelihood becomes

$$\pi(y \mid x, \gamma) \propto \exp\left(-\frac{1}{2}(y - \mathsf{A}_\gamma x)^{\mathsf{T}} \Sigma^{-1} (y - \mathsf{A}_\gamma x)\right),$$

and additional information concerning γ, for example, bound constraints, can be included via a hyperprior density.

The procedure just outlined can be applied to many problems arising from adaptive optics imaging in astronomy [52]; while the uncertainty in the model is more complex than in the explanatory example above, the approach remains the same.

Conditionally Gaussian hypermodels: Gaussian prior models are often criticized for being a too restricted class, not being able to adequately represent prior beliefs concerning, for example, the sparsity or piecewise smoothness of the solution. The range of qualitative features that can be expressed with normal densities can be considerably expanded by considering *conditionally Gaussian* families instead. As an example, consider the problem of finding a sparse image from linearly blurred noisy copy of it. The likelihood model in this case may be written as in (12). To set up an appropriate prior, consider a conditionally Gaussian prior

$$\begin{aligned} \pi_{\text{prior}}(x \mid \theta) &\propto \left(\frac{1}{\theta_1 \cdots \theta_N}\right)^{1/2} \exp\left(-\frac{1}{2}\sum_{j=1}^{N} \frac{x_j^2}{\theta_j}\right) \\ &= \exp\left(-\frac{1}{2}\sum_{j=1}^{N}\left[\frac{x_j^2}{\theta_j} + \log\theta_j\right]\right). \end{aligned} \tag{13}$$

If $\theta_j = \theta_0 =$ constant, we obtain the standard white noise prior which cannot be expected to favor sparse solutions. On the other hand, since θ_j is the variance of the pixel X_j, sparse images correspond to vectors θ with most of the components close to zero. Since we do not know a priori which of the variances should significantly differ from zero, when choosing a stochastic model for θ, it is reasonable to select a hyperprior that favors sparsity without actually specifying the location of the outliers. Two distributions that are particularly well suited for this are the *gamma distribution*,

$$\theta_j \sim \text{Gamma}(k, \theta_0), \quad k, \theta_0 > 0, \quad \pi(\theta_j) = \theta_j^{k-1} \exp\left(-\frac{\theta_j}{\theta_0}\right),$$

and the *inverse gamma distribution*,

$$\theta_j \sim \text{InvGamma}(k, \theta_0), \quad k, \theta_0 > 0, \quad \pi(\theta_j) = \theta_j^{-k-1} \exp\left(-\frac{\theta_0}{\theta_j}\right).$$

The parameters k and θ_0 are referred to as the shape and the scaling, respectively. The inverse gamma distribution corresponds to assuming that the *precision*, defined as $1/\theta_j$, is distributed according to the gamma distribution $\text{Gamma}(k, 1/\theta_0)$. The computational price of introducing hyperparameters is that instead of one image x, we need to estimate the image x and its variance image θ. Fortunately, for conditionally Gaussian families, there are efficient algorithms for computing these estimates, which will be discussed in the section concerning algorithms.

The hyperprior based on the gamma distribution, in turn, contains parameters (k and θ_0) to be determined. Nothing prevents us from defining another layer of hyperpriors concerning these values. It should be noted that in hierarchical models, the selection of the parameters higher up in the hierarchy tends to have less direct effect on the parameters of primary interest. Since this last statement has not been formally proved to be true, it should be considered as a piece of computational folklore.

Conditionally Gaussian hypermodels have been successfully applied in machine learning [66], in electromagnetic brain activity mapping [16], and in imaging applications for restoring blocky images [15]. Recently, their use in compressed sensing has been proposed [40].

4 Numerical Methods and Case Examples

The solution of an imaging inverse problem in the statistical framework is the posterior probability density. Because this format of the solution is not practical for most applications, it is common to summarize the distribution in one or a few images. This leads to the challenging problem of exploring the posterior distributions and finding single estimators supported by the distribution.

Estimators

In this section, we review some of the commonly used estimators and subsequently discuss some of the popular algorithms suggested in the literature to compute the corresponding estimates.

Prelude: Least Squares and Tikhonov Regularization

In the case where the forward model is linear, the problem of estimating an image from a degraded, noisy recording is equivalent in a determinist setting to looking for a solution of a linear system of equations of the form

$$\mathsf{A}x = y, \tag{14}$$

where the right-hand side is corrupt by noise. When A is not a square matrix and/or it is ill conditioned, one needs to specify what a "solution" means. The most straightforward way is to specify it as a least squares solution.

There is a large body of literature, and a wealth of numerical algorithms, for the solution of large-scale least squares problems arising from problems similar to imaging applications (see, e.g., [9]). Since dimensionality alone makes these problems computationally very demanding, they may require an unreasonable amount of computer memory and operations unless a compact representation of the matrix A can be exploited. Many of the available algorithms make additional assumptions about either the underlying image or the structure of the forward model regardless of whether there is a good justification.

In a determinist setting, the entries of the least squares solution of (14) with a right-hand side corrupted by noise are not necessarily in the gray-scale range of the image pixels. Moreover, the inherent ill conditioning of the problem, which varies with the imaging modality and the conditions under which the observations were collected, usually requires requlariza tion, see, for example, [4, 33, 34, 41]. A standard regularization method is to replace the original ill-posed least squares problem by a nearby well-posed problem by introducing a penalty term to avoid that the computed solution is dominated by amplified noise components, reducing the problem to minimizing a functional of the form

$$T(x) = \|\mathsf{A}x - y\|^2 + \alpha J(x), \tag{15}$$

where $J(x)$ is the penalty functional and $\alpha > 0$ is the regularization parameter. The minimizer of the functional (15) is the *Tikhonov regularized solution*. The type of additional information used in the design of the penalty term may include upper bounds on the norm of the solution or of its derivatives, nonnegative constraints for its entries, or bounds on some of the components. Often, expressing characteristics that are expected of the sought image in qualitative terms is neither new nor difficult: the translation of these beliefs into mathematical terms and their implementation is a more challenging step.

Maximum Likelihood and Maximum A Posteriori

We begin with the discussion of the maximum likelihood estimator in the framework of non-Bayesian statistics and denote by x a deterministic parameter determining the likelihood distribution of the data, modeled as a random variable. Let $\hat{x} = \hat{x}(y)$ denote an estimator of x, based on the observations y. Obviously, $\hat{x}$ is also a random variable, because of its dependency on the stochastic observations y; moreover, it is an *unbiased estimator* if

$$\mathsf{E}\{\hat{x}(y)\} = x,$$

that is, if, in the average, it returns the exact value. The covariance matrix C of an unbiased estimator therefore measures the statistical variation around the true value,

$$\mathsf{C}_{j,k} = \mathsf{E}\left\{(\hat{x}_j - x_j)(\hat{x}_k - x_k)\right\},$$

thus the name mean square error. Evidently, the smaller the mean square error, for example, in the sense of quadratic forms, the higher the expected fidelity of the estimator. The Fisher information matrix (9) gives a lower bound for the covariance matrix of all unbiased estimators. Assuming that J is invertible, the *Cramér–Rao lower bound* states that for an unbiased estimator,

$$\mathsf{J}^{-1} \leq \mathsf{C}$$

in the sense of quadratic forms, that is, for any vector

$$u^{\mathsf{T}}\mathsf{J}^{-1}u \leq u^{\mathsf{T}}\mathsf{C}u.$$

An estimator is called *efficient* if the error covariance reaches the Cramér–Rao bound.

The maximum likelihood estimator $\hat{x}_{\mathrm{ML}}(y)$ is the maximizer of the function $x \mapsto f(x; y)$, and in practice, it is found by locating the zero(s) of the score,

$$\nabla_x \log f(x; y) = 0 \Rightarrow x = \hat{x}_{\mathrm{ML}}(y).$$

Notice that in the non-Bayesian context, likelihood refers solely to the likelihood of the observations y, and the maximum likelihood estimation is a way to choose the underlying parametric model so that the observations become as likely as possible.

The popularity of the maximum likelihood estimator, in addition to being an intuitively obvious choice, stems from the fact that it is asymptotically efficient estimator in the sense that when the number of independent observations of the data increases, the covariance of the estimator converges toward the inverse of the Fisher information matrix, assuming that it exists. More precisely, assuming a sequence $y^1, y^2, \ldots$ of independent observations and defining $\hat{x}^n = \hat{x}\left(y^1, \ldots, y^n\right)$ as

$$\hat{x}^n = \mathrm{argmax}\left\{\frac{1}{n}\sum_{j=1}^{n} f\left(x, y^j\right)\right\},$$

asymptotically the probability distribution of $\hat{x}^n$ approaches a Gaussian distribution with mean x and covariance J^{-1}.

The assumption of the regularity of the Fisher information matrix limits the use of the ML estimator in imaging applications. To understand this claim, consider the simple case of linear forward model and additive Gaussian noise,

$$Y = \mathsf{A}x + E, \quad E \sim \mathcal{N}(0, \Sigma).$$

The likelihood function in this case is

$$f(x;y) = \left(\frac{1}{2\pi|\Sigma|}\right)^{1/N} \exp\left(-\frac{1}{2}(y-\mathsf{A}x)^{\mathsf{T}}\Sigma^{-1}(y-\mathsf{A}x)\right),$$

from which it is obvious that by formula (11),

$$\mathsf{J} = \mathsf{A}^{\mathsf{T}}\Sigma^{-1}\mathsf{A}.$$

In the simplest imaging problems such as of denoising, the invertibility of J is not an issue. However, in more realistic and challenging applications such as deblurring, the ill conditioning of A renders J singular, and the Cramér–Rao bound becomes meaningless. It is not uncommon to regularize the information matrix by adding a diagonal weight to it which, from the Bayesian viewpoint, is tantamount to adding prior information but in a rather uncontrolled manner.

For further reading of mathematical methods in estimation theory, we refer to [17, 46, 50].

We consider the maximum likelihood estimator in the context of regularization and Bayesian statistics. In the case of a Gaussian additive noise observation model, under the assumption that the noise at each pixel is independent of the signal and that the forward map is linear, $F(x) = \mathsf{A}x$, the likelihood (8) is of the form

$$\pi(y \mid x) \propto \exp\left(-\frac{1}{2}\|\mathsf{D}(\mathsf{A}x-y)\|^2\right),$$

where Σ is the noise covariance matrix and $\mathsf{D}^{\mathsf{T}}\mathsf{D} = \Sigma^{-1}$ is the Cholesky decomposition of its inverse. The maximizer of the likelihood function is the solution of the minimization problem

$$x_{\mathrm{ML}} = \operatorname{argmin}\left\{\|\mathsf{D}(\mathsf{A}x-y)\|^2\right\},$$

which, in turn, is the least squares solution of the linear system

$$\mathsf{D}\mathsf{A}x = \mathsf{D}y.$$

Thus, we can reinterpret least squares solutions as maximum likelihood estimates under an additive, independent Gaussian error model. Within the statistical framework, the maximum likelihood estimator is defined analogously for any error model which admits a maximizer for the likelihood, but in the general case, the computation of the minimizer cannot be reduced to the solution of a linear least squares problem.

In a statistical framework, the addition of a penalty terms to keep the solution of the least squares problem from becoming dominated by amplified noise components is tantamount to using a prior to augment the likelihood. If the observation model is linear, the prior and the likelihood are both Gaussian,

$$\pi_{\text{prior}}(x) \propto \exp\left(-\frac{1}{2}x^{\mathsf{T}}\Gamma^{-1}x\right),$$

and the noise is independent of the signal, the corresponding posterior is of the form

$$\pi(x \mid y) \propto \exp\left(-\frac{1}{2}\left(\|\mathsf{D}(\mathsf{A}x - y)\|^2 + \|\mathsf{R}x\|^2\right)\right),$$

where R satisfies $\mathsf{R}^{\mathsf{T}}\mathsf{R} = \Gamma^{-1}$, so typically it is the Cholesky factor of Γ^{-1} or alternatively, $\mathsf{R} = \Gamma^{-1/2}$.

The maximizer of the posterior density, or the maximum a posteriori (MAP) estimate, is the minimizer of the negative exponent, hence the solution of the minimization problem

$$\begin{aligned} x_{\text{MAP}} &= \text{argmin}\{\|\mathsf{D}(\mathsf{A}x - y)\|^2 + \|\mathsf{R}x\|^2\} \\ &= \text{argmin}\left\{\left\|\begin{bmatrix}\mathsf{DA}\\ \mathsf{R}\end{bmatrix}x - \begin{bmatrix}\mathsf{D}y\\ 0\end{bmatrix}\right\|^2\right\}, \end{aligned}$$

or, equivalently, the Tikhonov solution (15) with penalty $J(x) = \|\mathsf{R}x\|^2$ and regularization parameter $\alpha = 1$. Again, it is important to note that the direct correspondence between the Tikhonov regularization and the MAP estimate only holds for linear observation models and Gaussian likelihood and prior. The fact that the MAP estimate in this case is the least squares solution of the linear system

$$\begin{bmatrix}\mathsf{DA}\\ \mathsf{R}\end{bmatrix}x = \begin{bmatrix}\mathsf{D}y\\ 0\end{bmatrix} \tag{16}$$

is a big incentive to stay with Gaussian likelihood and Gaussian priors as long as possible.

As in the case of the ML estimate, the definition of MAP estimate is independent of the form of the posterior, hence applied also to non-Gaussian, nonindependent noise models, with the caveat that in the general case, the search for a maximizer of the posterior may require much more sophisticated optimization tools.

Conditional Means

The recasting in statistical terms of imaging problems effectively shifts the interest from the image itself to its probability density. The ML and MAP estimators discussed in the previous section suffer from the limitations, which come from summarizing an entire distribution with one realization. The ML estimator is known to suffer from instabilities due to the typical ill conditioning of the forward map in imaging problems, and it will not be discussed further here. The computed MAP estimate, on the other hand, may correspond to an isolated spike in the probability density away from the bulk of the mass of the density, and its computation may suffer from numerical complications. Furthermore, a conceptually more serious

limitation is the fact that MAP estimators do not carry information about the statistical dispersion of the distribution. A tight posterior density suggests that any ensemble of images which are in statistical agreement with the data and the given prior show little variability; hence, any realization from that ensemble can be thought of as very representative of the entire family. A wide posterior, on the other hand, suggests that there is a rather varied family of images that are in agreement with the data and the prior, hence lowering the representative power of any individual realization.

In the case where either the likelihood or the prior is not Gaussian, the mean of the posterior density, often referred to as conditional mean (CM) or posterior mean, may be a better choice because it is the estimator with least variance (see [3, 41]). Observe, however, that in the fully Gaussian case, the MAP and CM estimate coincides.

The CM estimate is, by definition,

$$x_{\mathrm{CM}} = \int_{\mathbb{R}^N} x\pi(x \mid y)dx,$$

while the a posteriori covariance matrix is

$$\Gamma_{\mathrm{CM}} = \int_{\mathbb{R}^N} (x - x_{\mathrm{CM}})(x - x_{\mathrm{CM}})^{\mathsf{T}}\pi(x \mid y)dx,$$

hence requiring the evaluation of the high-dimensional integrals. When the integrals have no closed form solution, as is the case for many imaging problems where, for example, the a priori information contains bounds on pixel values, a numerical approximation of the integral must be used to estimate x_{CM} and Γ_{CM}. The large dimensionality of the parameter space, which easily is of the order of hundreds of thousands when x represents an image, rules out the use of standard numerical quadratures, leaving Monte Carlo integration the only currently known feasible alternative.

The conceptual simplicity of Monte Carlo integration, which estimates the integral value as the average of a large sample of the integrand evaluated over the support of the integration, requires a way of generating a large sample from the posterior density. The generation of a sample from a given distribution is a well-known problem in statistical inference, which has inspired families of sampling schemes generically referred to as Markov chain Monte Carlo (MCMC) methods, which will be discussed in section "Markov Chain Monte Carlo Sampling."

Once a representative sample from the posterior has been generated, the CM estimate is approximately the sample mean. By definition, the CM estimate must be near the bulk of the density, although it is not necessarily a highly probable point. In fact, for multimodal distributions, the CM estimate may fall between the modes of the density and even belong to a subset of $\mathbb{R}^N$ with probability zero, although such a situation is rather easy to detect. There is evidence, however, that in some imaging applications the CM estimate is more stable than the MAP estimate; see

[23]. While the robustness of the CM estimate does not compensate for the lack of information about the width of the posterior, the possibility of estimating the posterior covariance matrix via sampling is an argument for the sampling approach, since the sample can also be used to estimate the posterior width.

Algorithms

The various estimators based on the posterior distribution are simple to define, but the actual computation may be a major challenge. In the case of Gaussian likelihood and prior, combined with linear forward map, the MAP and CM estimates coincide and an explicit formula exists. If the problem is very high dimensional, even this case may be computationally challenging. Before going to specific algorithms, we review the linear Gaussian theory.

The starting point is the linear additive model

$$Y = \mathsf{A}X + E, \quad X \sim \mathcal{N}(0, \Gamma), \quad E \sim \mathcal{N}(0, \Sigma).$$

Here, we assume that the mean of X and the noise E both vanish, an assumption that is easy to remove. Above, X and E need not be mutually independent, and we may postulate that they are jointly Gaussian and the cross-correlation matrix

$$\mathsf{C} = \mathsf{E}\left\{XE^{\mathsf{T}}\right\} \in \mathbb{R}^{N\times M}$$

may not vanish. The joint probability distribution of X and Y is also Gaussian, with zero mean and variance

$$\begin{aligned}\mathsf{E}\left\{\begin{bmatrix}X\\Y\end{bmatrix}\begin{bmatrix}X^{\mathsf{T}} & Y^{\mathsf{T}}\end{bmatrix}\right\} &= \mathsf{E}\left\{\begin{bmatrix}XX^{\mathsf{T}} & X(\mathsf{A}X+E)^{\mathsf{T}}\\ (\mathsf{A}X+E)X^{\mathsf{T}} & (\mathsf{A}X+E)(\mathsf{A}X+E)^{\mathsf{T}}\end{bmatrix}\right\}\\ &= \begin{bmatrix}\Gamma & \Gamma\mathsf{A}^{\mathsf{T}}+\mathsf{C}\\ \mathsf{A}\Gamma+\mathsf{C}^{\mathsf{T}} & \mathsf{A}\Gamma\mathsf{A}^{\mathsf{T}}+\Sigma\end{bmatrix}.\end{aligned}$$

Let $\mathsf{L} \in \mathbb{R}^{(N+M)\times(N+M)}$ denote the inverse of the above matrix, assuming that it exists, and write a partitioning of it in blocks according to the dimensions N and M,

$$\mathsf{L} = \begin{bmatrix}\Gamma & \Gamma\mathsf{A}^{\mathsf{T}}+\mathsf{C}\\ \mathsf{A}\Gamma+\mathsf{C}^{\mathsf{T}} & \mathsf{A}\Gamma\mathsf{A}^{\mathsf{T}}+\Sigma\end{bmatrix}^{-1} = \begin{bmatrix}\mathsf{L}_{11} & \mathsf{L}_{12}\\ \mathsf{L}_{21} & \mathsf{L}_{22}\end{bmatrix}.$$

With this notation, the joint probability distribution of X and Y is

$$\pi(x, y) \propto \exp\left(-\frac{1}{2}(x^{\mathsf{T}}\mathsf{L}_{11}x + x^{\mathsf{T}}\mathsf{L}_{12}y + y^{\mathsf{T}}\mathsf{L}_{21}x + y^{\mathsf{T}}\mathsf{L}_{22}y)\right).$$

To find the posterior density, one completes the square in the exponent with respect to x,

$$\pi(x \mid y) \propto \exp\left(-\frac{1}{2}\left(x - \mathsf{L}_{11}^{-1}\mathsf{L}_{12}y\right)^{\mathsf{T}}\mathsf{L}_{11}\left(x - \mathsf{L}_{11}^{-1}\mathsf{L}_{12}y\right)\right),$$

where terms independent of x that contribute only to the normalization are left out. Therefore,

$$X \mid Y = y \sim \mathcal{N}\left(\mathsf{L}_{11}^{-1}\mathsf{L}_{12}y, \mathsf{L}_{11}^{-1}\right).$$

Finally, we need to express the matrix blocks L_{ij} in terms of the matrices of the model. The expressions follow from the classical matrix theory of Schur complements [24]: We have

$$\mathsf{L}_{11}^{-1} = \Gamma - \left(\Gamma\mathsf{A}^{\mathsf{T}} + \mathsf{C}\right)\left(\mathsf{A}\Gamma\mathsf{A}^{\mathsf{T}} + \Sigma\right)^{-1}\left(\mathsf{A}\Gamma + \mathsf{C}^{\mathsf{T}}\right), \tag{17}$$

and

$$\mathsf{L}_{11}^{-1}\mathsf{L}_{12}y = \left(\Gamma\mathsf{A}^{\mathsf{T}} + \mathsf{C}\right)\left(\mathsf{A}\Gamma\mathsf{A}^{\mathsf{T}} + \Sigma\right)^{-1}y. \tag{18}$$

Although a closed form solution, to evaluate the expression (18) for the posterior mean may require iterative solvers.

When the image and the noise are mutually independent, implying that $\mathsf{C} = 0$, we find a frequently encountered form of the MAP estimate arising from writing the Gaussian posterior density directly by using Bayes' formula, that is,

$$\begin{aligned}\pi(x \mid y) &\propto \pi_{\text{prior}}(x)\pi(y \mid x)\\ &\propto \exp\left(-\frac{1}{2}x^{\mathsf{T}}\Gamma^{-1}x - \frac{1}{2}(y - \mathsf{A}x)^{\mathsf{T}}\Sigma^{-1}(y - \mathsf{A}x)\right),\end{aligned}$$

and so the MAP estimate, and simultaneously the posterior mean estimate, is the maximizer of the above expression, or, equivalently, the minimizer of the quadratic functional

$$H(x) = (y - \mathsf{A}x)^{\mathsf{T}}\Sigma^{-1}(y - \mathsf{A}x) + x^{\mathsf{T}}\Gamma^{-1}x.$$

By substituting the factorizations

$$\Sigma^{-1} = \mathsf{D}^{\mathsf{T}}\mathsf{D}, \quad \Gamma^{-1} = \mathsf{R}^{\mathsf{T}}\mathsf{R},$$

the minimization problem becomes the previously discussed standard least squares problem of minimizing

$$H(x) = \|\mathsf{D}(y - \mathsf{A}x)\|^2 + \|\mathsf{R}x\|^2, \tag{19}$$

leading to the least squares problem (16). Whether one should use this formula or (18) depends on the application and, in particular, on the sparsity properties of the covariance matrices and their inverses.

Iterative Linear Least Squares Solvers

The computation of the ML or MAP estimate under the Gaussian additive linear noise model and, in the latter case, with a Gaussian prior, amounts to the solution of system of linear equations (14), (16), or (18) in the least squares sense. Since the dimensions of the problem are proportional to the number of pixels in the image except when the observation model has a particular structure or sparsity properties which can be exploited to reduce the memory allocation, solution by direct methods is unfeasible, hence making in general the iterative solvers the methods of choice.

Among the iterative methods specifically designed for the solution of least squares problems, the LSQR version with shifts [55, 56] of the Conjugate Gradient for Least Squares (CGLS) method originally proposed in [37] combines robustness and numerical efficiency. CGLS-type iterative methods have been designed to solve the system $\mathsf{A}x = y$, minimize $\|\mathsf{A}x - y\|^2$, or minimize $\|\mathsf{A}x - y\|^2 + \delta\|x\|^2$, where the matrix A may be square or rectangular – either overdetermined or underdetermined – and may have any rank. The matrix A does not need to be stored, but instead its action is represented by a routine for computing matrix–vector products of the forms $v \mapsto \mathsf{A}v$ and $u \mapsto \mathsf{A}^{\mathsf{T}}u$.

Minimizing the expression (19) may be transformed in a standard form by writing it as

$$\min\left\{\|\mathsf{D}\left(y - \mathsf{A}\mathsf{R}^{-1}w\right)\|^2 + \|w\|^2\right\}, \quad w = \mathsf{R}x$$

In practice, the matrix R^{-1} should not be computed, unless it is trivial to obtain. Rather, R^{-1} acts as a preconditioner, and its action should be implemented together with the action of the matrix A as a routine called from the iterative linear solver. The interpretation of the action of the prior as a preconditioner has led to the concept of prior conditioner; see [12, 14] for details.

Nonlinear Maximization

In the more general case where either the observation model is nonlinear or the likelihood and prior are non-Gaussian, the computation of the ML and MAP estimates requires the solution of a maximization problem. Maximizers of nonlinear functions can be found by quasi-Newton methods with global convergence strategy. Since Newton-type methods proceed by solving a sequence of linearized problems whose dimensions are proportional to the size of the image, iterative linear solvers are typically used for the solution of the linear subproblem [20, 43]. In imaging applications, it is not uncommon that the a priori information includes nonnegativity constraints on the pixel values or bounds on their range. In these cases, the computation of the MAP estimate amounts to a constrained maximization problem and may be very challenging. Algorithms for maximization problems with nonnegativity constraints arising in imaging applications based on the projected

gradient have been proposed in the literature; see [2] and references therein. We shall not review Newton-based methods here, since usually the fine points are related to the particular applications at hand and not so much to the statistical description of the problem. Instead, we review some algorithms that stem directly from the statistical setting of the problem and are therefore different from the methods used in regularized deterministic literature.

EM Algorithm

The MAP estimator is the maximizer of the posterior density $\pi(x \mid y)$, or, equivalently, the maximizer the logarithm of it,

$$L(x \mid y) = \log \pi(x \mid y) = \log \pi(y \mid x) + \log \pi_{\text{prior}}(x) + \text{constant},$$

where the simplest form of Bayes' rule was used to represent the posterior density as a product of the likelihood and the prior. However, note that above, the vector x may represent the unknown of primary interest, or if hierarchical models are used, the model parameters related to the likelihood and/or prior may be included in it.

The *expectation–maximization* algorithm is a method developed originally for maximizing the likelihood function and later extended to the Bayesian setting to maximize the posterior density, in a situation where part of the data is "missing." While in many statistical application the concept of missing data appears natural, for example, when incomplete census data or patient data are discussed, in imaging applications, this concept is a rather arbitrary and to some extent artificial. However, during the years, EM has found its way to numerous imaging applications, partly because it often leads to algorithms that are easy to implement. Early versions of the imaging algorithms with counting data such as the Richardson–Lucy iteration [49, 57], popular in astronomical imaging, were independently derived. Later, similar EM-based algorithms were rederived in the context of medical imaging [29, 36, 60]. Although EM algorithms are discussed in more detail elsewhere in this book, we include a brief discussion here in order to put EM in the context of general statistical imaging formalism.

As pointed out above, in imaging problems, data is not missing: Data, *per definitionem*, is what one is able to observe and register. Therefore, the starting point of the EM algorithm in image applications is to augment the actual data y by *fictitious*, nonexistent data z that would make the problem significantly easier to handle.

Consider the statistical inference problem of estimating a random variable X based on an observed realization of Y, denoted by $Y = y = y_{\text{obs}}$. We assume the existence of a third random variable Z and postulate that the joint probability density of these three variables is available and is denoted by $\pi(x, y, z)$. The EM algorithm consists of the following steps:

1. Initialize $x = x^0$ and set $k = 0$.
2. *E-step*: Define the probability distribution, or a fictitious likelihood density,

$$\pi^k(z) = \pi\left(z \mid x^k, y\right) \propto \pi\left(x^k, y, z\right), \quad y = y_{\text{obs}},$$

and calculate the integral

$$Q^k(x) = \int L(x \mid y, z)\pi^k(z)dz, \quad L(x \mid y, z) = \log(\pi(x \mid y, z)). \tag{20}$$

3. *M-step*: Update x^k by defining

$$x^{k+1} = \text{argmax}\, Q^k(x). \tag{21}$$

4. If a given convergence criterion is satisfied, exit; otherwise, increase k by one and repeat from step 2 until convergence.

The E-step above can be interpreted as computing the expectation of the real-valued random variable $\log(\pi(x, y, Z))$, x and y fixed, with respect to a conditional measure of Z conditioned on $X = x^j$ and $Y = y = y_{\text{obs}}$, hence the name expectation step.

The use of the EM algorithm is often advocated on the basis of the convergence proof given in [19]. Unfortunately, the result is often erroneously quoted as an automatic guarantee of convergence, without verifying the required hypotheses. The validity of the convergence is further obfuscated by the error in the proof (see [70]), and in fact, counterexamples of lack of convergence are well known [10, 69]. We point out that as far as convergence is concerned, global convergence of quasi-Newton algorithm is well established, and compared to the EM algorithm, the algorithm is often more effective [20].

As the concept of missing data is not well defined in general, we outline the use of the EM algorithm in an example that is meaningful in imaging applications.

SPECT imaging: The example discussed here follows the article [29]. Consider the SPECT image formation problem, where the two-dimensional object is divided in N pixels, each one emitting photons that are recorded through collimators by M photon counting devices. If x_j is the expected number of photons emitted by the jth pixel, the photon count at ith photon counter, denoted by Y_i, is an integer-valued random variable and can be modeled by a Poisson process,

$$Y_i \sim \text{Poisson}\left(\sum_{j=1}^{M} a_{ij}x_j\right) = \text{Poisson}((\mathsf{A}x)_i),$$

the variables Y_i being mutually independent and the matrix elements a_{ij} of $\mathsf{A} \in \mathbb{R}^{M\times N}$ being known. We assume that X, the stochastic extension of the unknown vector $x \in \mathbb{R}^N$, is a priori distributed according to a certain probability distribution,

$$X \sim \pi_{\text{prior}}(x) \propto \exp(-V(x)).$$

To apply the EM algorithm, we need to decide how to define the "missing data." Photon counter devices detect the emitted photons added over the line of sight; evidently, the problem would be more tractable if we knew the number of emitted photons from each pixel separately. Therefore, we define a fictitious measurement,

$$Z_{ij} \sim \text{Poisson}(a_{ij} x_j),$$

and posit that these variables are mutually independent. Obviously, after the measurement $Y = y$, we have

$$\sum_{j=1}^{N} Z_{ij} = y_i. \tag{22}$$

To perform the E-step, assuming that x^k is given, consider first the conditional density $\pi^k(z) = \pi\left(z \mid x^k, y\right)$.

A basic result from probability theory states that if N independent random variables Λ_j are a priori Poisson distributed with respective means μ_j, and in addition

$$\sum_{j=1}^{N} \Lambda_j = K,$$

then, a posteriori, the variables Λ_j conditioned on the above data are binomially distributed,

$$\Lambda_j \mid \left(\sum_{j=1}^{N} \Lambda_j = K \right) \sim \text{Binom}\left(K, \frac{\mu_j}{\sum_{j=1}^{N} \mu_j} \right).$$

In particular, the conditional expectation of Λ_j is

$$\mathsf{E}\left\{ \Lambda_j \mid \sum_{j=1}^{N} \Lambda_j = K \right\} = K \frac{\mu_j}{\sum_{j=1}^{N} \mu_j}.$$

We therefore conclude that the conditional density $\pi^k(z)$ is a product of binomial distributions of Z_{ij} with a priori means $\mu_j = a_{ij} x_j^k$, $\sum_{j=1}^{N} \mu_j = (\mathsf{A}x^k)_i$, and $K = y_i$, so in particular,

$$\mathsf{E}\left\{ Z_{ij} \mid \sum_{j=1}^{N} Z_{ij} = y_i \right\} = \int z_{ij} \pi^k(z) dz = y_i \frac{a_{ij} x_j^k}{(\mathsf{A}x^k)_i} \stackrel{\text{def}}{=} z_{ij}^k. \tag{23}$$

Furthermore, by Bayes' theorem,

$$\pi(x \mid y, z) = \pi(x \mid z) = \pi(z \mid x)\pi_{\text{prior}}(x),$$

where we used the fact that the true observations y add no information on x that would not be included in z, we have, by definition of the Poisson likelihood and the prior,

$$L(x \mid y, z) = \sum_{ij} \big(z_{ij} \log(a_{ij}x_j) - a_{ij}x_j\big) - V(x) + \text{constant},$$

and therefore, up to an additive constant, we have

$$Q^k(x) = \sum_{ij} \Big(z_{ij}^k \log(a_{ij}x_j) - a_{ij}x_j\Big) - V(x),$$

where z_{ij}^k is defined in (23). This completes the E-step.

The M-step requires the minimization of $Q^k(x)$ given above. Assuming that V is differentiable, the minimizer should satisfy

$$\frac{1}{x_\ell}\sum_{i=1}^m z_{i\ell}^k - \sum_{i=1}^m a_{i\ell} - \frac{\partial V}{\partial x_\ell}(x) = 0.$$

How complicated it is to find a solution to this condition depends on the prior contribution V and may require an internal Newton iteration. In [29], an approximate "one-step late" (OSL) algorithm was suggested, which is tantamount to a fixed-point iteration: Initiating with $\tilde{x}^0 = x^k$, an update scheme $\tilde{x}^t \to \tilde{x}^{t+1}$ is given by

$$\tilde{x}_\ell^{t+1} = \frac{\sum_{i=1}^m z_{i\ell}^k}{\sum_{i=1}^m a_{i\ell} + \frac{\partial V}{\partial x_\ell}(\tilde{x}^t)},$$

and this step is repeated until a convergence criterion is satisfied at some $t = t*$. Finally, the M-step is completed by updating $x^{k+1} = \tilde{x}^{t*}$.

The EM algorithm has been applied to other imaging problems such as blind deconvolution problem [44] and PET imaging [36, 71].

Markov Chain Monte Carlo Sampling

In Bayesian statistical imaging, the real solution of the imaging problem is the posterior density of the image interpreted as a multivariate random variable. If a closed form of the posterior is either unavailable or not suitable for the tasks at hand, the alternative is to resort to exploring the density by generating a representative sample from it. Markov chain Monte Carlo (MCMC) samplers yield samples from a target distribution by moving from a point in a chain to the next by the transition

rule which characterizes the specific algorithm. MCMC sampling algorithms are usually subdivided into those which are variants of the Metropolis–Hastings (MH) algorithm or the Gibbs sampler. While the foundations of the MH algorithm were laid first [25, 35, 51], Gibbs samplers have sometimes the appeal of being more straightforward to implement.

The basic idea of Monte Carlo integration is rather simple. Assume that $\pi(x)$ is a probability density in $\mathbb{R}^N$, and let $\{X^1, X^2, X^3, \ldots\}$ denote a stochastic process, where the random variables X^i are independent and identically distributed, $X^i \sim \pi(x)$. The central limit theorem asserts that for any measurable $f : \mathbb{R}^N \to \mathbb{R}$,

$$\frac{1}{n}\sum_{i=1}^{n} f(X^i) \overset{n\to\infty}{\longrightarrow} \int_{\mathbb{R}^N} f(x)\pi(x)dx \quad \text{almost certainly,} \tag{24}$$

and moreover, the convergence takes place asymptotically with the rate $1/\sqrt{n}$, independently of the dimension N. The difficulty is to find a computationally efficient way of drawing independently from a given distribution π. Indeed, when N is large, it may be even difficult to decide where the numerical support of the density is. In MCMC methods, instead of producing an independent chain, the idea is to produce a Markov process $\{X^i\}$ with the property that π is the equilibrium distribution. It can be shown (see [53, 61, 65]) that with rather mild assumptions (irreducibility, aperiodicity), the limit (24) holds, due to the law of large numbers.

In applications to imaging, the computational burden associated with MCMC methods has become proverbial and is often presented as the main obstacle to the use of Bayesian method in imaging. It is easy to imagine that sampling random variable with hundreds of thousands of components will require a large amount of computer resources and that collecting and storing a large number of images will require much more time than estimating a single one. On the other hand, since an ensemble of images from a distribution carries a lot of additional information which cannot be included in single-point estimates, it seems unreasonable to rate methods simply according to computational speed. That said, since collecting a well-mixed, representative sample poses several challenges, in the description of the Gibbs sampling and Metropolis–Hastings algorithms, we will point out references to variants which can improve the independence and mixing of the ensemble; see [30–32].

In its first prominent appearance in the imaging arena [26], the Gibbs sampler was presented as part of a stochastic relaxation algorithm to efficiently compute MAP estimates. The systematic or fully conditional Gibbs sampler algorithm proceeds as follows [61].

Let $\pi(x)$ be a probability density defined on $\mathbb{R}^N$, denoted by $\pi(x) = \pi(x_1, \ldots, x_N)$, $x \in \mathbb{R}^N$ to underline that it is the joint density of the components of X. Furthermore, denote by $\pi(x_j \mid x_{-j})$ the conditional density of the jth component x_j given all the other components, collected in the vector $x_{-j} \in \mathbb{R}^{N-1}$. Let x^1 be the initial element of the Markov chain. Assuming that we are at a point x^i in the chain, we need a rule stating how to proceed to the next point x^{i+1}, i.e.,

we need to describe the updating method of proceeding from the current element x^i to x^{i+1}. This is done by updating sequentially each component as follows.

Fully conditional Gibbs sampling update: Given x^i, compute the next element x^{i+1} by the following algorithm:

$$\begin{aligned}
&\text{draw} \quad x_1^{i+1} \quad \text{from} \quad \pi\left(x_1 \mid x_{-1}^i\right);\\
&\text{draw} \quad x_2^{i+1} \quad \text{from} \quad \pi\left(x_2 \mid x_1^{i+1}, x_3^i, \ldots, x_N^i\right);\\
&\text{draw} \quad x_3^{i+1} \quad \text{from} \quad \pi\left(x_3 \mid x_1^{i+1}, x_2^{i+1}, x_4^i \ldots, x_N^i\right);\\
&\qquad \vdots\\
&\text{draw} \quad x_N^{i+1} \quad \text{from} \quad \pi\left(x_N \mid x_{-N}^{i+1}\right).
\end{aligned}$$

In imaging applications, this Gibbs sampler may be impractical because of the large number of components of the random variable to be updated to generate a new element of the chain. In addition, if some of the components are correlated, updating them independently may slow down the chain to explore the full support of the distribution, due to slow movement at each step. The correlation among components can be addressed by updating blocks of correlated components together, although this will imply that the draws must be from multivariate instead of univariate conditional densities.

It follows naturally from the updating scheme that the speed at which the chain will reach equilibrium is strongly dependent on how the system of coordinate axes relates to the most prominent correlation directions. A modification of the Gibbs sampler that can ameliorate the problems caused by correlated components performs a linear transformation of the random variable using correlation information. Without going into details, we refer to [48, 58, 61] for different variants of Gibbs sampler.

The strategy behind the Metropolis–Hastings samplers is to generate a chain with the target density as equilibrium distribution by constructing at each step the transition probability function from the current $X^i = x$ to next realization of X^{i+1} in the chain in the following way. Given an initial transition probability function $q(x, x')$ with $X^i = x$, x' drawn from $q(x, x')$ is a *proposal* for the value of X^{i+1}. Upon acceptance of $X^{i+1} = x'$, which occurs with probability $\alpha(x, x')$, defined by

$$\alpha(x, x') = \min\left\{\frac{\pi(x')q(x', x)}{\pi(x, x')}, 1\right\}, \quad \pi(x)q(x, x') > 0.$$

We add it to the chain; otherwise, we reject the proposed value and we set $X^{i+1} = x$. In the latter case, the chain did not move and the value x is replicated in the chain. The transition probability $p(x, x')$ of the Markov chain thus defined is

$$p(x, x') = q(x, x')\alpha(x, x'),$$

while the probability to stay put is

$$1 - \int_{\mathbb{R}^N} q(x, y)\alpha(x, y)dy.$$

This construction guarantees that the transition probability satisfies the detailed balance equation $\pi(x)p(x, x') = \pi(x')p(x', x)$, from which it follows that, for reasonable choices of the function q, $\pi(x)$ is the equilibrium distribution of the chain.

This algorithm is particularly convenient when the target distribution $\pi(x)$ is a posterior. In fact, since the only way in which π enters is via the ratio of its values at two points, it is sufficient to compute the density modulo a proportionality constant, which is how we usually define the posterior. Specific variants of the MH algorithm correspond to different choices of $q(x, x')$; in the original formulation [51], a symmetric proposal, for example, a random walk, was used, so that $q(x, x') = q(x', x)$, implying that

$$\alpha(x, x') = \min\{\pi(x')/\pi(x), 1\},$$

while the general formulation above is due to Hastings [35]. An overview of the different possible choices for $q(x, x')$ can be found in [65].

A number of hybrid sampling schemes which combine different chains or use MH variants to draw from the conditional densities inside Gibbs samplers have been proposed in the literature; see [48, 61] and references therein. Since the design of efficient MCMC samplers must address the specific characteristics of the target distribution, it is to be expected that as the use of densities becomes more pervasive in imaging, new hybrid MCMC scheme will be proposed.

The convergence of Monte Carlo integration based on MCMC methods is a key factor in deciding when to stop sampling. This is particularly pertinent in imaging applications, where the calculations needed for additions of a point to the chain may be quite time consuming. Due to the lack of a systematic way of translating theoretical convergence results of MCMC chains [7, 65] into pragmatic stopping rules, in practice, the issue is reduced to monitoring the behavior of the already collected sample.

As already pointed out, MCMC algorithms are not sampling independently from the posterior. When computing sample-based estimates for the posterior mean and covariance,

$$\hat{x}_{\mathrm{CM}} = \frac{1}{n}\sum_{j=1}^{n} x^j, \quad \hat{\Gamma}_{\mathrm{CM}} = \frac{1}{n}\sum_{j=1}^{n} \left(x^j - \hat{x}_{\mathrm{CM}}\right)\left(x^j - \hat{x}_{\mathrm{CM}}\right)^{\mathsf{T}}.$$

A crucial question is how accurately these estimates approximate the posterior mean and covariance. The answer depends on the sample size n and the sampling strategy itself. Ideally, if the sample vectors x^j are realizations of independent identically

distributed random variables, the approximations converge with the asymptotic rate $1/\sqrt{n}$, in agreement with the central limit theorem. In practice, however, the MCMC sampling produces sample points that are mutually correlated, and the convergence is slower.

The convergence of the chain can be investigated using the *autocovariance function* (ACF) of the sample [27, 64]. Assume that we are primarily interested in estimating a real-valued function $f : \mathbb{R}^N \to \mathbb{R}$ of the unknown, and we have generated an MCMC sample or a realization $\{x^1, \ldots, x^n\}$ of a stationary stochastic process $\{X^1, \ldots, X^n\}$. The random variables X^j are equally distributed, their distribution being the posterior distribution $\pi(x)$ of a random variable X. The estimation of the mean quantity $f(X)$ can be done by calculating

$$\hat{\mu} = \frac{1}{n} \sum_{j=1}^{n} f(x^j),$$

while the theoretical mean of $f(X)$ is

$$\mu = \mathsf{E}\{f(X)\} = \int f(x)\pi(x)dx.$$

Each sample yields a slightly different value for $\hat{\mu}$, which is itself a realization of the random variable F defined as

$$F = \frac{1}{n} \sum_{j=1}^{n} f(X^j).$$

The problem is now how to estimate the variance of F, which gives us an indication of how well the computed realization approximates the mean. The identical distribution of the random variables X^j implies that

$$\mathsf{E}\{F\} = \frac{1}{n} \sum_{j=1}^{n} \underbrace{\mathsf{E}\{f(X^j)\}}_{=\mu} = \mu,$$

while the variance of F, which we want to estimate starting from the available realization of the by stochastic process, is

$$\mathrm{var}(F) = \mathsf{E}\{F^2\} - \mu^2.$$

To this end, we need to introduce some definitions and notations.

We define the autocovariance function of the stochastic process $f(X^j)$ with lag $k \geq 0$ to be

$$C(k) = \mathsf{E}\{f(X^j) f(X^{j+k})\} - \mu^2$$

which, if the process is stationary, is independent of j. The normalized ACF is defined as

$$c(k) = \frac{C(k)}{C(0)}.$$

The ACF can be estimated from an available realization as follows

$$\hat{C}(k) = \frac{1}{n-k} \sum_{j=1}^{n-k} f(x^j) f(x^{j+k}) - \hat{\mu}^2. \tag{25}$$

It follows from the definition of F that

$$\mathsf{E}\left\{F^2\right\} = \frac{1}{n^2} \sum_{i,j=1}^{n} \mathsf{E}\left\{f(X^i) f(X^j)\right\}.$$

Let us now focus on the random matrix $\left[f(X^i) f(X^j)\right]_{i,j=1}^{n}$. The formula above takes its expectation and subsequently computes the average of its entries. By stationarity, the expectation is a symmetric Toeplitz matrix; hence, its diagonal entries are all equal to

$$\mathsf{E}\left\{f(X^i) f(X^i)\right\} = C(0) + \mu^2,$$

while the kth subdiagonal entries are all equal to

$$\mathsf{E}\left\{f(X^i) f(X^{i+k})\right\} = C(k) + \mu^2.$$

This observation provides us with a simple way to perform the summation by accounting for the elements along the diagonals, leading to the formula

$$\mathsf{E}\left\{F^2\right\} = \frac{1}{n^2} \left(nC(0) + 2 \sum_{k=1}^{n-1} (n-k) C(k) \right) + \mu^2,$$

from which it follows that the variance of F is

$$\operatorname{var}(F) = \frac{1}{n} \left(C(0) + 2 \sum_{k=1}^{n-1} \left(1 - \frac{k}{n}\right) C(k) \right).$$

If we assume that the ACF is negligible when $k > n_0$, for some n_0 significantly smaller than the sample size n, we may use the approximation

$$\operatorname{var}(F) \approx \frac{1}{n} \left(C(0) + 2 \sum_{k=1}^{n_0} C(k) \right) = \frac{C(0)}{n} \tau,$$

where

$$\tau = 1 + 2\sum_{k=1}^{n_0} c(k). \tag{26}$$

If we account fully for all contributions,

$$\tau = 1 + 2\sum_{k=1}^{n-1} \left(1 - \frac{k}{n}\right) c(k), \tag{27}$$

which is the Cesàro mean of the normalized ACFs or low-pass filtered mean with the triangular filter. The quantity τ is called the *integrated autocorrelation time* (IACT) and can be interpreted as the time that it takes for our MCMC to produce an independent sample. If the convergence rate for independence samplers is $1/\sqrt{n}$, the convergence rate for the MCMC sampler is $1/\sqrt{n/\tau}$. If the variables X_j are independent, then $\tau = 1$, and the result is exactly what we would expect from the central limit theorem, because in this case, $C(0) = n\,\mathrm{var}(f(X))$.

The estimate of τ requires an estimate for the normalized ACF, which can be obtained with the formula (25), and a value for n_0 to use in formula (26). In the choice of n_0, it is important to remember that $\hat{C}(k)$ is a realization of a random sequence $C(k)$, which in practice contains noise. Some practical rules for choosing n_0 are suggested in [27].

In [27], it is shown that since the sequence

$$\gamma(k) = c(2k) + c(2k+1), \quad k = 0, 1, 2, \ldots$$

is *strictly positive*, *strictly decreasing*, and *strictly convex*, that is,

$$\gamma(k) > 0, \quad \gamma(k+1) < \gamma(k), \quad \gamma(k+1) < \frac{1}{2}(\gamma(k) + \gamma(k+2)),$$

when the sample-based estimated sequence,

$$\hat{\gamma}(k) = \hat{c}(2k) + \hat{c}(2k+1), \quad k = 0, 1, 2, \ldots$$

fails to be so, this is an indication that the contribution is predominantly coming from noise; hence, it is wise to stop summing the terms to estimate τ. Geyer proposes three initial sequence estimators, in the following order:

1. Initial positive sequence estimator (IPSE): Choose n_0 to be the largest integer for which the sequence remains positive,

$$n_0 = n_{\mathrm{IPSE}} = \max\{k \mid \gamma(k) > 0\}.$$

2. Initial monotone sequence estimator (IMSE): Choose n_0 to be the largest integer for which the sequence remains positive and monotone,

$$n_0 = n_{\text{IMSE}} = \max\{k \mid \gamma(k) > 0,\ \gamma(k) < \gamma(k-1)\}.$$

3. Initial convex sequence estimator (ICSE): Choose n_0 to be the largest integer for which the sequence remains positive, monotone, and convex,

$$n_0 = n_{\text{ICSE}} = \max\left\{k \mid \gamma(k) > 0,\ \gamma(k) < \gamma(k-1),\ \gamma(k-1) < \frac{1}{2}(\gamma(k) + \gamma(k-2))\right\}.$$

From the proof in [27], it is obvious that also the sequence $\{c(k)\}$ itself must be positive and decreasing. Therefore, to find n_0 for IPSE or IMSE, there is no need for passing to the sequence $\{\gamma(k)\}$. As for ICSE, again from the proof in the cited article, it is also clear that the sequence

$$\eta(k) = c(2k+1) + c(2k+2), \quad k = 0, 1, 2, \ldots$$

too, is positive, monotonous, and convex. Therefore, to check the condition for ICSE, it might be advisable to form both sequences $\{\gamma(k)\}$ and $\{\eta(k)\}$ and set n_{ICSE} equal to the maximum index for which both $\gamma(k)$ and $\eta(k)$ remain strictly convex.

Summarizing a practical rule, using for instance, the IMSE, to compute τ is:

1. Estimate the ACF sequence $\hat{C}(k)$ from the sample by formula (25) and normalize it by $\hat{C}(0)$ to obtain $\hat{c}(k)$.
2. Find n_0 equal to the largest integer for which the sequence $\hat{c}(0), \hat{c}(1), \ldots, \hat{c}(n_0)$ remains positive and strictly decreasing. Notice that the computation of ACFs can be stopped when such an n_0 is reached.
3. Calculate the estimate for the IACT τ,

$$\tau = 1 + 2\sum_{k=1}^{n_0}\left(1 - \frac{k}{n}\right)c(k) \approx 1 + 2\sum_{k=1}^{n_0} c(k). \tag{28}$$

Notice that if n is not much larger than n_0, the sample is too small.

The accuracy of the approximation of μ by $\hat{\mu}$ is often expressed, with some degree of imprecision, by writing an estimate

$$\mu = \hat{\mu} \pm 2\left(\frac{C(0)}{n}\tau\right)^{1/2}$$

with the 95 % belief. This interpretation is based on the fact that, with a probability of about 95 %, the values of a Gaussian random variable are within ± 2 STD from

the mean. Such an approximate claim is justified when n is large, in which case the random variable F is asymptotically Gaussian by the central limit theorem.

Statistical Approach: What Is the Gain?

Statistical methods are often pitted against deterministic ones, and the true gain of the approach is sometimes lost, especially if the statistical methods are used only to produce single estimates. Indeed, it is not uncommon that the statistical framework is seen simply as an alternative way of explaining regularization. Another criticism of statistical methods concerns the computation times. While there is no doubt that computing a posterior mean using MCMC methods is more computationally intensive than resorting to optimization-based estimators, it is also obvious that a comparison in these terms does not make much sense, since a sample contains enormously more information of the underlying distribution than an estimate of its mode.

To emphasize what there is to be gained when using the statistical approach, we consider some algorithms that have been found useful and are based on the interpretation images as random variables.

Beyond the Traditional Concept of Noise

The range of interpretation of the concept of noise in imaging is usually very restricted, almost exclusively referring to uncertainties in observed data due to exogenous sources. In the context of deterministic regularization, the noise model is almost always additive, in agreement with the paradigm that only acknowledges noise as the difference between a "true" and "noisy" data, giving no consideration to its statistical properties. Already the proper noise modeling of counting data clearly demonstrates the shortcomings of such models. The Bayesian – or subjective – use of probability as an expression of uncertainty allows to extend the concept of noise to encompass a much richer terrain of phenomena, including shortcomings in the forward model, prior, or noise statistics itself.

To demonstrate the possibilities of the Bayesian modeling, consider an example where it is assumed that a forward model with additive noise,

$$y = F(x) + e. \tag{29}$$

which describes, to the best of our knowledge, as completely as possible, the interdependency of the data y and the unknown. We refer to it as the *detailed model*. Here, the noise e is thought to be exogenous, and its statistical properties are known.

Assume further that the detailed model is computationally too complex to be used with the imaging algorithms and the application at hand for one or several of the following reasons. The dimensionality of the image x may be too high for the model to be practical; the model may contain details such as boundary conditions that need to be simplified in practice; the deblurring kernel may be non-separable, while in practice, a fast algorithm for separable kernels may exist. To

accommodate these difficulties, a simpler model is constructed. Let z be possibly a simpler representation of x, obtained, for example, via a projection to a coarser grid, and let f denote the corresponding forward map. It is a common procedure to write a simplified model of the form

$$y = f(z) + e, \tag{30}$$

which, however, may not explain the data as well as the detailed model (29). To properly account for the errors added by the model reduction, we should write instead

$$\begin{aligned} y &= F(x) + e = f(z) + [F(x) - f(z)] + e \\ &= f(z) + \varepsilon(x, z) + e, \quad \varepsilon(x, z) = F(x) - f(z), \end{aligned} \tag{31}$$

where the term $\varepsilon(x, z)$ is referred to as *modeling error*.

In the framework of deterministic imaging, modeling errors pose unsurmountable problems because they depend on both the unknown image x and its reduced counterpart z. A common way to address errors coming from model reduction is to artificially increase the variance of the noise included in the reduced model until it masks the modeling error. Such an approach introduces a statistical structure in the noise that does not correspond to the modeling error and may easily waste several orders of magnitude of the accuracy of the data. On the other hand, neglecting the error introduced by model reduction may lead to overly optimistic estimates of the performance of algorithms. The very questionable procedure of testing algorithms with data simulated with the same forward map used for the inversion is referred to as *inverse crime* [42]. Inverse criminals, who tacitly assume that $\varepsilon(x, z) = 0$, should not be surprised if the unrealistically good results obtained from simulated data are not robust when using real data.

While modeling error often is neglected also in the statistical framework, its statistical properties can be described in terms of the prior. Consider the stochastic extension of $\varepsilon(x, z)$,

$$\tilde{E} = \varepsilon(X, Z),$$

where X and Z are the stochastic extensions of x and z, respectively. Since, unlike an exogenous noise term, the modeling error is not independent of the unknowns Z and X, the likelihood and the prior cannot be described separately, but instead must be specified together.

To illustrate how ubiquitous modeling error is, consider the following example.

Boundary clutter and image truncation: Consider a denoising/deblurring example of the type encountered in astronomy, microscopy, and image processing. Let $u : \mathbb{R}^2 \to \mathbb{R}$ be a continuous two-dimensional model of a scenery that is recorded through an out-of-focus device. The noiseless model for the continuous problem is a convolution integral,

$$v(r) = \int_{\mathbb{R}^2} a(r-s)u(s)ds, \quad r \in \mathbb{R}^2,$$

the convolution kernel $a(r-s)$ describing the point spread of the device. We assume that $r \mapsto a(r)$ decays rapidly enough to justify an approximation as a compactly supported function.

Let $Q \subset \mathbb{R}^2$ define a bounded *field of view*. We consider the following imaging problem: *Given a noisy version of the blurred image v over the field of view Q, estimate the underlying image u over the field of view Q.*

Assume that a sufficiently fine discretization of Q into N pixels is given, and denote by $r_i \in Q$ the center of the ith pixel. Assume further that the point spread function a is negligibly small outside a disc D of radius $\delta > 0$. By selecting an *extended field of view* Q' such that

$$Q + D = \{s \in \mathbb{R}^2 \mid s = r + r',\ r \in Q,\ r' \in D\} \subset Q',$$

we may restrict the domain of integration in the definition of the convolution integral

$$v(r_i) = \int_{\mathbb{R}^2} a(r_i - s)u(s)ds \approx \int_{Q'} a(r_i - s)u(s)ds.$$

After discretizing Q' into N' pixels p_j with center points s_j, N of which are within the field of view, coinciding with R^J we can restate the problem in the form

$$\begin{aligned} v(s_i) &\approx \int_{Q'} a(s_i - s)u(s)ds \approx \sum_{j=1}^{N'} |p_j| a(s_i - s_j)u(s_j) \\ &= a_{ij}u(s_j), \quad a_{ij} = |p_j| a(s_i - s_j), \quad 1 \le i \le N. \end{aligned}$$

After accounting for the contribution of exogenous noise at each recorded pixel, we arrive at the complete discrete model

$$y = \mathsf{A}'x + e, \quad \mathsf{A}' \in \mathbb{R}^{N \times N'}, \tag{32}$$

where $x_j = u(s_j)$ and y_i represent the noisy observation of $v(s_i)$. If the pixelization is fine enough, we may consider this model to be a good approximation of the continuous problem.

A word of caution is in order when using this model, because the right-hand side depends not only on pixels within the field of view, where we want to estimate the underlying image, but also on pixels in the frame $C = Q' \setminus Q$ around it. The vector x is therefore partitioned into two vectors, where the first one, denoted by $z \in \mathbb{R}^N$, contains values in the pixels within the field of view, and the second one, $\zeta \in \mathbb{R}^K$, $K = N' - N$, consists of values of pixels in the frame. After suitably rearranging the indices, we may write x in the form

$$x = \begin{bmatrix} z \\ \\ \zeta \end{bmatrix} \begin{matrix} \mathbb{R}^N \\ \in \times \\ \mathbb{R}^K \end{matrix},$$

and, after partitioning the matrix A' accordingly,

$$\mathsf{A}' = \begin{bmatrix} \mathsf{A} & \mathsf{B} \end{bmatrix} \in \mathbb{R}^{N \times N} \times \mathbb{R}^{N \times K},$$

we can rewrite the model (32) in the form

$$y = \mathsf{A}z + \mathsf{B}\zeta + e = \mathsf{A}z + \varepsilon + e,$$

where the modeling errors are collected in second term ε, which we will refer to as *boundary clutter*. It is well known that ignoring the contribution to the recorded image coming from and beyond the boundary may cause severe artifacts in the estimation of the image x within the field of view. In a determinist framework, the boundary clutter term is often compensated for by extending the image outside the field of view in a manner believed to be closest to the actual image behavior. Periodic extension or extensions obtained by reflecting the image symmetrically or antisymmetrically are quite popular in the literature, because they will significantly simplify the computations; details on such an approach can be found, for example, in [21].

Consider a Gaussian prior and a Gaussian likelihood,

$$X \sim \mathcal{N}(0, \Gamma), \quad E \sim \mathcal{N}(0, \Sigma_{\text{noise}}),$$

and partition the prior covariance matrix according to the partitioning of x,

$$\Gamma \in \begin{bmatrix} \Gamma_{11} & \Gamma_{12} \\ \Gamma_{21} & \Gamma_{22} \end{bmatrix}, \quad \Gamma_{11} \in \mathbb{R}^{N \times N}, \, \Gamma_{12} = \Gamma_{21}^{\mathsf{T}} \in \mathbb{R}^{N \times K}, \, \Gamma_{22} \in \mathbb{R}^{K \times K}.$$

The covariance matrix of the total noise term, which also includes the boundary clutter $\tilde{E}$, is

$$\mathsf{E}\left\{\left(\tilde{E} + E\right)\left(\tilde{E} + E\right)^{\mathsf{T}}\right\} = \mathsf{B}\Gamma_{22}\mathsf{B}^{\mathsf{T}} + \Sigma_{\text{noise}} = \Sigma$$

and the cross covariance of the image within the field of view and the noise is

$$\mathsf{C} = \mathsf{E}\left\{Z\left(\tilde{E} + E\right)^{\mathsf{T}}\right\} = \Gamma_{12}\mathsf{B}^{\mathsf{T}}.$$

The posterior distribution of the vector Z conditioned on $Y = y$ now follows from (17) and (18). The posterior mean is

$$z_{\mathrm{CM}} = \left(\Gamma_{11}\mathsf{A} + \Gamma_{12}\mathsf{B}^{\mathsf{T}}\right)\left(\mathsf{A}\Gamma_{11}\mathsf{A}^{\mathsf{T}} + \mathsf{B}\Gamma_{22}\mathsf{B}^{\mathsf{T}} + \Sigma_{\mathrm{noise}}\right)^{-1} y,$$

and the posterior covariance is

$$\Gamma_{\mathrm{post}} = \Gamma_{11} - \left(\Gamma_{11}\mathsf{A} + \Gamma_{12}\mathsf{B}^{\mathsf{T}}\right)\left(\mathsf{A}\Gamma_{11}\mathsf{A}^{\mathsf{T}} + \mathsf{B}\Gamma_{22}\mathsf{B}^{\mathsf{T}} + \Sigma_{\mathrm{noise}}\right)^{-1}\left(\Gamma_{11}\mathsf{A} + \Gamma_{12}\mathsf{B}^{\mathsf{T}}\right)^{\mathsf{T}}.$$

A computationally efficient and robust algorithm for computing the conditional mean is proposed in [13]. For further applications of the modeling error approach in imaging, see [1, 38, 47].

Sparsity and Hypermodels

The problem of reconstructing sparse images or more generally images that can be represented as sparse linear combinations of prescribed basis images using data consisting of few measurements has recently received a lot of attention and has become a central issue in compressed sensing [11]. Bayesian hypermodels provide a very natural framework for deriving algorithms for sparse reconstruction.

Consider a linear model with additive Gaussian noise, the likelihood being given by (12) and a conditionally Gaussian prior (13) with hyperparameter θ. As explained in section "Adding Layers: Hierarchical Models," if we select the hyperprior $\pi_{\mathrm{hyper}}(\theta)$ in such a way that it favors solutions with variances Θ_j close to zero except for only few outliers, the overall prior for (X, Θ) will be biased toward sparse solutions. Two hyperpriors well suited for sparse solutions are the gamma and the inverse gamma hyperpriors. For the sake of definiteness, consider the inverse gamma hyperprior with mutually independent components,

$$\pi_{\mathrm{hyper}}(\theta_j) = \theta_j^{-k-1}\exp\left(-\frac{\theta_0}{\theta_j}\right) = \exp\left(-\frac{\theta_0}{\theta_j} - (k+1)\log\theta_j\right).$$

Then the posterior distribution for the pair (X, Θ) is of the form

$$\pi(x, \theta \mid y) \propto \exp\left(-\frac{1}{2}(y - \mathsf{A}x)^{\mathsf{T}}\Sigma^{-1}(y - \mathsf{A}x) - \frac{1}{2}x^{\mathsf{T}}\mathsf{D}_\theta^{-1}x - \sum_{j=1}^{N} V(\theta_j)\right)$$

where

$$V(\theta_j) = \frac{\theta_0}{\theta_j} + \left(k + \frac{3}{2}\right)\log\theta_j, \quad \mathsf{D}_\theta = \mathrm{diag}(\theta) \in \mathbb{R}^{N\times N}.$$

An estimate for (X, Θ) can be found by maximizing $\pi(x, \theta \mid y)$ with respect to the pair (x, θ) using, for example, a quasi-Newton optimization scheme. Alternatively, the following two algorithms that make use of the special form of the expression above can also be used.

In the articles [66, 67] on Bayesian machine learning, the starting point is the observation that the posterior density $x \mapsto \pi(x, \theta \mid y)$ is Gaussian and therefore it is possible to integrate it explicitly with respect to x. It can be shown, after some tedious but straightforward algebraic manipulations, that the marginal posterior distribution is

$$\begin{aligned}\pi(\theta \mid y) &= \int_{\mathbb{R}^N} \pi(x, \theta \mid y) dx \\ &\propto \left(\frac{1}{\det(\mathsf{M}_\theta)}\right)^{1/2} \exp\left(-\sum_{j=1}^{N} V(\theta_j) + \frac{1}{2}\tilde{y}^{\mathsf{T}} \mathsf{M}_\theta^{-1} \tilde{y}\right),\end{aligned}$$

where

$$\mathsf{M}_\theta = \mathsf{A}^{\mathsf{T}} \Sigma^{-1} \mathsf{A} + \mathsf{D}_\theta^{-1}, \quad \tilde{y} = \mathsf{A}^{\mathsf{T}} \Sigma^{-1} y.$$

The *most probable* estimate or the *maximum evidence* estimator $\hat{\theta}$ of Θ is, by definition, the maximizer of the above marginal, or equivalently, the maximizer of its logarithm,

$$L(\theta) = -\frac{1}{2} \log\left(\det(\mathsf{M}_\theta)\right) - \sum_{j=1}^{N} V(\theta_j) + \frac{1}{2}\tilde{y}^{\mathsf{T}} \mathsf{M}_\theta^{-1} \tilde{y}$$

which must satisfy

$$\frac{\partial L}{\partial \theta_j} = 0, \quad 1 \leq j \leq N.$$

It turns out that, although the computation of the determinant may in general be a challenge, its derivatives can be expressed in a formally simple form. To this end separate the element depending on θ_j from D_θ^{-1}, writing

$$\mathsf{D}_\theta^{-1} = \frac{1}{\theta_j} e_j e_j^{\mathsf{T}} + \mathsf{D}_{\theta'}^{\dagger},$$

where e_j is the jth coordinate unit vector, θ' is the vector θ with the jth element replaced by a zero and "†" denotes the pseudo-inverse. Then

$$\begin{aligned}\mathsf{M}_\theta &= \mathsf{A}^{\mathsf{T}} \Sigma^{-1} \mathsf{A} + \mathsf{D}_{\theta'}^{\dagger} + \frac{1}{\theta_j} e_j e_j^{\mathsf{T}} = \mathsf{M}_{\theta'} + \frac{1}{\theta_j} e_j e_j^{\mathsf{T}} \\ &= \mathsf{M}_{\theta'} \left(\mathsf{I} + \frac{1}{\theta_j} q e_j^{\mathsf{T}}\right), \quad q = \mathsf{M}_{\theta'}^{-1} e_j.\end{aligned} \tag{33}$$

It follows from the properties of the determinant that

$$\det(\mathsf{M}_\theta) = \det\left(\mathsf{I} + \frac{1}{\theta_j} q e_j^{\mathsf{T}}\right) \det(\mathsf{M}_{\theta'}) = \left(1 + \frac{q_j}{\theta_j}\right) \det(\mathsf{M}_{\theta'}),$$

where $q_j = e_j^{\mathsf{T}} q$. After expressing the inverse of M_θ in the expression of $L(\theta)$ via the Sherman–Morrison–Woodbury formula [28] as

$$\mathsf{M}_\theta^{-1} = \mathsf{M}_{\theta'}^{-1} - \frac{1}{\theta_j + q_j} q q^{\mathsf{T}},$$

we find that the function $L(\theta)$ can be written as

$$L(\theta) = \frac{1}{2} \log\left(1 + \frac{q_j}{\theta_j}\right) - V(\theta_j) + \frac{1}{2} \frac{(q^{\mathsf{T}} \tilde{y})^2}{\theta_j + q_j} + \text{terms that are independent of } \theta_j.$$

The computation of the derivative of $L(\theta)$ with respect to θ_j and its zeros is now straightforward, although not without challenges because reevaluation of the vector q may potentially be expensive. For details, we refer to the article [67].

After having found an estimate $\hat{\theta}$, an estimate for X can be obtained by observing that the conditional density $\pi(x \mid y, \hat{\theta})$ is Gaussian,

$$\pi(x \mid y, \hat{\theta}) \propto \exp\left(-\frac{1}{2}(y - \mathsf{A}x)^{\mathsf{T}} \Sigma^{-1} (y - \mathsf{A}x) - \frac{1}{2} x^{\mathsf{T}} \hat{\theta} x\right),$$

and an estimate for x is obtained by solving in the least squares sense the linear system

$$\begin{bmatrix} \Sigma^{-1/2}\mathsf{A} \\ \mathsf{D}_{\hat{\theta}}^{-1/2} \end{bmatrix} x = \begin{bmatrix} \Sigma^{-1/2} y \\ 0 \end{bmatrix}. \tag{34}$$

In imaging applications, this is a large-scale linear problem and typically, iterative solvers need to be employed [59].

A different approach leading to a fast algorithm of estimating the MAP estimate $(x, \theta)_{\text{MAP}}$ was suggested in [15]. The idea is to maximize the posterior distribution using an alternating iteration: Starting with an initial value $\theta = \theta^1$, $\ell = 1$, the iteration proceeds as follows:

1. Find $x^{\ell+1}$ that maximizes $x \mapsto L(x, \theta^\ell) = \log\big(\pi(x, \theta^\ell \mid y)\big)$.
2. Update $\theta^{\ell+1}$ by maximizing $\theta \mapsto L(x^{\ell+1}, \theta) = \log\big(\pi(x^{\ell+1}, \theta \mid y)\big)$.

The efficiency of this algorithm is based on the fact that for $\theta = \theta^\ell$ fixed, the maximization of $L(x, \theta^\ell)$ in the first step is tantamount to minimizing the quadratic expression

$$\frac{1}{2}\|\Sigma^{-1/2}(y - \mathsf{A}x)\|^2 + \frac{1}{2}\left\|\mathsf{D}_{\theta^\ell}^{-1/2}x\right\|^2,$$

the non-quadratic part being independent of x. Thus, step 1 only requires an (approximate) linear least squares solution of the system similar to (34). On the other hand, when $x = x^{\ell+1}$ is fixed, the minimizer of the second step is found as a zero of the gradient of the function $L(x^{\ell+1}, \theta)$ with respect to θ. This step, too, is straightforward, since the component equations are mutually independent,

$$\frac{\partial}{\partial\theta_j}L(x^{\ell+1}, \theta) = -\left(\frac{1}{2}\left(x_j^{\ell+1}\right)^2 + \theta_0\right)\frac{1}{\theta_j^2} + \left(k + \frac{3}{2}\right)\frac{1}{\theta_j} = 0,$$

leading to the explicit updating formula

$$\theta_j^{\ell+1} = \frac{1}{2k+3}\left(\left(x_j^{\ell+1}\right)^2 + 2\theta_0\right).$$

For details and performance of the method in image applications, we refer to [15].

5 Conclusion

This chapter gives an overview of statistical methods in imaging. Acknowledging that it would be impossible to give a comprehensive review of all statistical methods in imaging in a chapter, we have put the emphasis on the Bayesian approach, while making repeated forays in the frequentists' field. These editorial choices are reflected in the list of references, which only covers a portion of the large body of literature published on the topic. The use of statistical methods in subproblems of imaging science is much wider than presented here, extending, for example, from image segmentation to feature extraction, interpretation of functional MRI signals, and radar imaging.

Cross-References

▸ EM Algorithms
▸ Iterative Solution Methods
▸ Linear Inverse Problems
▸ Total Variation in Imaging

References

1. Arridge, S.R., Kaipio, J.P., Kolehmainen, V., Schweiger, M., Somersalo, E., Tarvainen, T., Vauhkonen, M.: Approximation errors and model reduction with an application in optical diffusion tomography. Inverse. Probl. **22**, 175–195 (2006)
2. Bardsley, J., Vogel, C.R.: A nonnegatively constrained convex programming method for image reconstruction. SIAM J. Sci. Comput. **25**, 1326–1343 (2004)
3. Bernardo, J.: Bayesian Theory. Wiley, Chichester (2000)
4. Bertero, M., Boccacci, P.: Introduction to Inverse Problems in Imaging. IOP, Bristol (1998)
5. Besag, J.: Spatial interaction and the statistical analysis of lattice systems. J. Stat. R. Soc. **36**, 192–236 (1974)
6. Besag, J.: On the statistical analysis of dirty pictures. J. R. Stat. Soc. B **48**, 259–302 (1986)
7. Besag, J., Green, P.: Spatial statistics and Bayesian computation. J. R. Stat. Soc. B **55**, 25–37 (1993)
8. Billingsley, P.: Probability and Measure, 3rd edn. Wiley, New York (1995)
9. Björck, Å.: Numerical Methods for Least Squares Problems. SIAM, Philadelphia (1996)
10. Boyles, R.A.: On the convergence of the EM algorithm. J. R. Stat. Soc. B **45**, 47–50 (1983)
11. Bruckstein, A.M., Donoho, D.L., Elad, M.: From sparse solutions of systems of equations to sparse modeling of signals and images. SIAM Rev. **51**, 34–81 (2009)
12. Calvetti, D.: Preconditioned iterative methods for linear discrete ill-posed problems from a Bayesian inversion perspective. J. Comput. Appl. Math. **198**, 378–395 (2007)
13. Calvetti, D., Somersalo, E.: Statistical compensation of boundary clutter in image deblurring. Inverse Probl. **21**, 1697–1714 (2005)
14. Calvetti, D., Somersalo, E.: Introduction to Bayesian Scientific Computing – Ten Lectures on Subjective Probability. Springer, Berlin (2007)
15. Calvetti, D., Somersalo, E.: Hypermodels in the Bayesian imaging framework. Inverse Probl. **24**, 034013 (2008)
16. Calvetti, D., Hakula, H., Pursiainen, S., Somersalo, E.: Conditionally Gaussian hypermodels for cerebral source localization. SIAM J. Imaging Sci. **2**, 879–909 (2009)
17. Cramér, H.: Mathematical Methods in Statistics. Princeton University Press, Princeton (1946)
18. De Finetti, B.: Theory of Probability, vol 1. Wiley, New York (1974)
19. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via EM algorithm. J. R. Stat. Soc. B **39**, 1–38 (1977)
20. Dennis, J.E., Schnabel, R.B.: Numerical Methods for Unconstrained Optimization and Nonlinear Equations. SIAM, Philadelphia (1996)
21. Donatelli, M., Martinelli, A., Serra-Capizzano, S.: Improved image deblurring with anti-reflective boundary conditions. Inverse Probl. **22**, 2035–2053 (2006)
22. Franklin, J.N.: Well-posed stochastic extension of ill-posed linear problem. J. Math. Anal. Appl. **31**, 682–856 (1970)
23. Fox, C., Nicholls, G.: Exact MAP states and expectations from perfect sampling: Greig, Porteous and Seheult revisited. AIP Conf. Proc. ISSU **568**, 252–263 (2001)
24. Gantmacher, F.R.: Matrix Theory. AMS, New York (1990)
25. Gelfand, A.E., Smith, A.F.M.: Sampling-based approaches to calculating marginal densities. J. Am. Stat. Assoc. **85**, 398–409 (1990)
26. Geman, S., Geman, D.: Stochastic relaxation, Gibbs distributions and Bayesian rerstoration of images. IEEE Trans. Pattern Anal. Mach. Intell. **6**, 721–741 (1984)
27. Geyer, C.: Practical Markov chain Monte Carlo. Stat. Sci. **7**, 473–511 (1992)
28. Golub, G., VanLoan, C.F.: Matrix Computations. Johns Hopkins University Press, London (1996)
29. Green, P.J.: Bayesian reconstructions from emission tomography data using modified EM algorithm. IEEE Trans. Med. Imaging **9**, 84–93 (1990)
30. Green, P.J., Mira, A.: Delayed rejection in reversible jump Metropolis-Hastings. Biometrika **88**, 1035–1053 (2001)

31. Haario, H., Saksman, E., Tamminen, J.: An adaptive Metropolis algorithm. Bernoulli **7**, 223–242 (2001)
32. Haario, H., Laine, M., Mira, A., Saksman, E.: DRAM: efficient adaptive MCMC. Stat. Comput. **16**, 339–354 (2006)
33. Hansen, P.C.: Rank-Deficient and Ill-Posed Inverse Problems. SIAM, Philadelphia (1998)
34. Hansen, P.C.: Discrete Inverse Problems. Insights and Algorithms. SIAM, Philadelphia (2010)
35. Hastings, W.K.: Monte Carlo sampling methods using Markov chains and their applications. Biometrika **57**, 97–109 (1970)
36. Herbert, T., Leahy, R.: A generalized EM algorithm for 3D Bayesian reconstruction from Poisson data using Gibbs priors. IEEE Trans. Med. Imaging **8**, 194–202 (1989)
37. Hestenes, M.R., Stiefel, E.: Methods of conjugate gradients for solving linear systems. J. Res. Natl. Bur. Stand. **49**, 409–436 (1952)
38. Huttunen, J.M.J., Kaipio, J.P.: Model reduction in state identification problems with an application to determination of thermal parameters. Appl. Numer. Math. **59**, 877–890 (2009)
39. Jeffreys, H.: An invariant form for the prior probability in estimation problem. Proc. R. Soc. Lond. A **186**, 453–461 (1946)
40. Ji, S., Carin, L.: Bayesian compressive sensing and projection optimization. In: Proceedings of 24th International Conference on Machine Learning, Cornvallis (2007)
41. Kaipio, J., Somersalo, E.: Statistical and Computational Inverse Problems. Springer, Berlin (2004)
42. Kaipio, J.P., Somersalo, E.: Statistical inverse problems: discretization, model reduction and inverse crimes. J. Comput. Appl. Math. **198**, 493–504 (2007)
43. Kelley, T.: Iterative Methods for Optimization. SIAM, Philadelphia (1999)
44. Lagendijk, R.L., Biemond, J.: Iterative Identification and Restoration of Images. Kluwer, Boston (1991)
45. Laksameethanasan, D., Brandt, S.S., Engelhardt, P., Renaud, O., Shorte, S.L.: A Bayesian reconstruction method for micro-rotation imaging in light microscopy. Microsc. Res. Tech. **71**, 158–167 (2007)
46. LeCam, L.: Asymptotic Methods in Statistical Decision Theory. Springer, New York (1986)
47. Lehikoinen, A., Finsterle, S., Voutilainen, A., Heikkinen, L.M., Vauhkonen, M., Kaipio, J.P.: Approximation errors and truncation of computational domains with application to geophysical tomography. Inverse Probl. Imaging **1**, 371–389 (2007)
48. Liu, J.S.: Monte Carlo Strategies in Scientific Computing. Springer, Berlin (2003)
49. Lucy, L.B.: An iterative technique for the rectification of observed distributions. Astron. J. **79**, 745–754 (1974)
50. Melsa, J.L., Cohn, D.L.: Decision and Estimation Theory. McGraw-Hill, New York (1978)
51. Metropolis, N., Rosenbluth, A.W., Teller, A.H., Teller, E.: Equations of state calculations by fast computing machines. J. Chem. Phys. **21**, 1087–1092 (1953)
52. Mugnier, L.M., Fusco, T., Conan, J.-L.: Mistral: a myopic edge-preserving image restoration method, with application to astronomical adptive-optics-corrected long-exposure images. J. Opt. Soc. Am. A **21**, 1841–1854 (2004)
53. Nummelin, E.: MC's for MCMC'ists. Int. Stat. Rev. **70**, 215–240 (2002)
54. Ollinger, J.M., Fessler, J.A.: Positron-emission tomography. IEEE Signal Proc. Mag. **14**, 43–55 (1997)
55. Paige, C.C., Saunders, M.A.: LSQR: an algorithm for sparse linear equations and sparse least squares. TOMS **8**, 43–71 (1982)
56. Paige, C.C., Saunders, M.A.: Algorithm 583; LSQR: sparse linear equations and least-squares problems. TOMS **8**, 195–209 (1982)
57. Richardson, H.W.: Bayesian-based iterative method of image restoration. J. Opt. Soc. Am. **62**, 55–59 (1972)
58. Robert, C.P., Casella, G.: Monte Carlo Statistical Methods. Springer, New York (2004)
59. Saad, Y.: Iterative Methods for Sparse Linear Systems. SIAM, Philadelphia (2003)
60. Shepp, L.A., Vardi, Y.: Maximum likelihood reconstruction in positron emission tomography. IEEE Trans. Med. Imaging **MI-1**, 113–122 (1982)

61. Smith, A.F.M., Roberts, R.O.: Bayesian computation via Gibbs sampler and related Markov chain Monte Carlo methods. J. R. Stat. Soc. B **55**, 3–23 (1993)
62. Snyder, D.L.: Random Point Processes. Wiley, New York (1975)
63. Starck, J.L., Pantin, E., Murtagh, F.: Deconvolution in astronomy: a review. Publ. Astron. Soc. Pac. **114**, 1051–1069 (2002)
64. Tan, S.M., Fox, C., Nicholls, G.K.: Lecture notes (unpublished), Chap 9. http://www.math.auckland.ac.nz/
65. Tierney, L.: Markov chains for exploring posterior distributions. Ann. Stat. **22**, 1701–1762 (1994)
66. Tipping, M.E.: Sparse Bayesian learning and the relevance vector machine. J. Mach. Learn. Res. **1**, 211–244 (2001)
67. Tipping, M.E., Faul, A.C.: Fast marginal likelihood maximisation for sparse Bayesian models. In: Proceedings of the 19th International Workshop on Artificial Intelligence and Statistics, Key West, 3–6 Jan 2003
68. Van Kempen, G.M.P., Van Vliet, L.J., Verveer, P.J.: A quantitative comparison of image restoration methods in confocal microscopy. J. Microsc. **185**, 354–365 (1997)
69. Wei, G.C.G., Tanner, M.A.: A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algoritms. J. Am. Stat. Assoc. **85**, 699–704 (1990)
70. Wu, J.: On the convergence properties of the EM algorithm. Ann. Stat. **11**, 95–103 (1983)
71. Zhou, J., Coatrieux, J.-L., Bousse, A., Shu, H., Luo, L.: A Bayesian MAP-EM algorithm for PET image reconstruction using wavelet transform. Trans. Nucl. Sci. **54**, 1660–1669 (2007)

Supervised Learning by Support Vector Machines

Gabriele Steidl

Contents

1 Introduction 1394
2 Historical Background 1396
3 Mathematical Modeling and Applications 1398
Linear Learning 1398
Nonlinear Learning 1410
4 Survey of Mathematical Analysis of Methods 1428
Reproducing Kernel Hilbert Spaces 1428
Quadratic Optimization 1436
Results from Generalization Theory 1440
5 Numerical Methods 1445
6 Conclusion 1448
Cross-References 1449
References 1449

Abstract

During the last two decades, support vector machine learning has become a very active field of research with a large amount of both sophisticated theoretical results and exciting real-world applications. This paper gives a brief introduction into the basic concepts of supervised support vector learning and touches some recent developments in this broad field.

G. Steidl (✉)
Department of Mathematics, University of Kaiserslautern, Kaiserslautern, Germany
e-mail: steidl@mathematik.uni-kl.de

© Springer Science+Business Media New York 2015

O. Scherzer (ed.), *Handbook of Mathematical Methods in Imaging*,
DOI 10.1007/978-1-4939-0790-8_22

1 Introduction

The desire to learn from examples is as old as mankind but has reached a new dimension with the invention of computers. This paper concentrates on learning by support vector machines (SVMs) which meanwhile deliver state-of-the-art performance in many real-world applications. However, it should be mentioned at the beginning that there exist many alternatives to SVMs ranging from classical k-nearest neighbor methods over trees and neural networks to other kernel-based methods. Overviews can be found, e.g., in the books of Mitchell [73], Hastie et al. [47], Duda et al. [34], and Bishop [9].

SVMs are a new-generation learning system based on various components Including:

- Statistical learning theory,
- Optimization theory (duality concept),
- Reproducing kernel Hilbert spaces (RKHSs),
- Efficient numerical algorithms.

This synthesis and their excellent performance in practice make SVM-like learning attractive for researchers from various fields. A non-exhaustive list of SVM applications includes *text categorization* (see Joachims [53] and Leopold and Kinderman [64]), *handwritten character recognition* (see LeCun et al. [62]), *texture and image classification* (see Chapelle et al. [23]), *protein homology detection* (see Jaakkola and Haussler [51]), *gene expression* (see Brown et al. [17]), *medical diagnostics* (see Strauss et al. [96]), and *pedestrian and face detection*, (see Osuna et al. [77] and Viola and Jones [110]). There exist various benchmark data sets for testing and comparing new learning algorithms and a good collection of books and tutorials on SVMs as those of Vapnik [105], Burges, [19], Cristianini and Shawe-Taylor [27], Herbrich [48], Schölkopf and Smola [87], and Steinwart and Christmann [93]. The first and latter ones contain a mathematically more rigerous treatment of statistical learning aspects. So-called least squares SVMs are handled in the book of Suykens et al. [98], and SVMs from the approximation theoretic point of view are considered in the book of Cucker and Zhou [29].

Let $\mathcal{X} \subset \mathbb{R}^d$, $\mathcal{Y} \subset \mathbb{R}^{\tilde{d}}$, where for simplicity only $\tilde{d} = 1$ is considered, and $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$. The aim of the following sections is to learn a target function $\mathcal{X} \to \mathcal{Y}$ from given training samples $Z := \{(x_1, y_1), \dots, (x_m, y_m)\} \subset \mathcal{Z}$. A distinction is made between classification and regression tasks. In *classification* $\mathcal{Y}$ is a discrete set, in general as large as the number of classes the samples belong to. Here binary classification with just two labels in $\mathcal{Y}$ was most extensively studied. An example where binary classification is useful is SPAM detection. Another example in medical diagnostics is given in Fig. 1. Here it should be mentioned that in many practical applications, the original input variables are pre-processed to transform them into a new useful space which is often easier to handle but preserves the necessary discriminatory information. This process is also known as *feature extraction*.

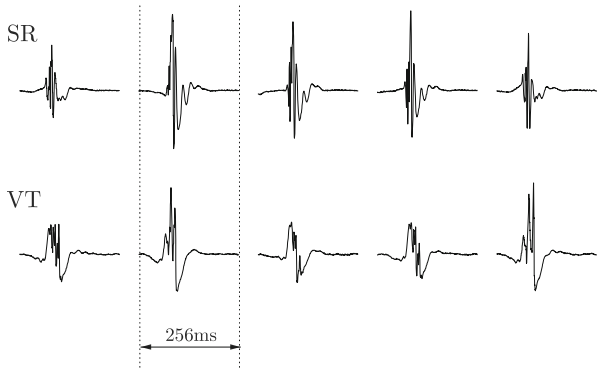


Fig. 1 Examples of physiological (SR) and pathological (VT) electrical heart activity curves measured by an implanted cardioverter-defibrillator. For the classification of such signals, see Strauss and Steidl [95]

In contrast to classification, *regression* aims at approximating the "whole" real-valued function from some function values, so that $\mathcal{Y}$ is not countable here. The above examples, as all problems considered in this paper, are from the area of *supervised learning*. This means that all input vectors come along with their corresponding target function values (labeled data). In contrast, *semi-supervised learning* makes use of labeled and unlabeled Data, and in *unsupervised learning* labeled data are not available, so that one can only exploit the input vectors x_i. The latter methods can be applied for example to discover groups of similar exemplars within the data (clustering), to determine the distribution of the data within the input space (density estimation), or to perform projections of data from high-dimensional spaces to lower-dimensional spaces. There are also learning models which involve more complex interactions between the learner and the environment. An example is *reinforcement learning* which is concerned with the problem of finding suitable actions in a given situation in order to maximize the reward. In contrast to supervised learning, reinforcement learning does not start from given optimal (labeled) outputs but must instead find them by a process of trial and error. For reinforcement learning, the reader may consult the book of Sutton and Barto [97].

Learning models can also differ in the way in which the training data are generated and presented to the learner. For example, a distinction can be made between *batch learning*, where all the data are given at the start of the training phase, and *online learning*, where the learner receives one example at a time and updates the hypothesis function in response to each new example.

This paper is organized as follows: An overview of the historical background is given in Sect. 2. Section 3 contains an introduction into classical SVM methods. It starts with linear methods for (binary) support vector classification and regression and considers also linear least squares classification/regression. Then the kernel trick is explained and used to transfer the linear models into so-called feature spaces which results in nonlinear learning methods. Some other models related to SVM as

well as multi-class classification and multitask learning are addressed at the end of the section. Section 4 provides some mathematical background concerning RKHSs and quadratic optimization. The last subsection sketches very briefly some results in statistical learning theory. Numerical methods to make the classical SVMs efficient in particular for large data sets are presented in Sect. 5. The paper ends with some conclusions in Sect. 6.

2 Historical Background

Modern learning theory has a long and interesting history going back as far as Gauss and Legendre but got its enormous impetus from the advent of computing machines. In the 1930s revolutionary changes took place in understanding the principles of inductive inference from a philosophical perspective, e.g., by Popper, and from the point of view of statistical theory, e.g., by Kolmogorov, Glivenko, and Cantelli, and applied statistics, e.g., by Fisher. A good overview over the leading ideas and developments in this time can be found in the comments and bibliographical remarks of Vapnik's book, Vapnik [105]. The starting point of statistical learning theory which considers the task of minimizing a risk functional based on empirical data dates back to the 1960s. Support vector machines, including their RKHS interpretation, were only discovered in the 1990s and led to an explosion in applications and theoretical analysis.

Let us start with the problem of linear regression which is much older than linear classification. The method of least squares was first published by Legendre [63]. It was considered as a statistical procedure by Gauss [43], who claimed, to the annoyance of Legendre but in accordance with most historians, to have applied this method since 1795. The original least squares approach finds for given points $x_i \in \mathbb{R}^d$ and corresponding $y_i \in \mathbb{R}$, $i = 1, \ldots, m$ a hyperplane $f(x) = \langle w, x\rangle + b$ having minimal least squares distance from the points (x_i, y_i):

$$\sum_{i=1}^{m} (\langle w, x_i\rangle + b - y_i)^2 \;\to\; \min_{w,b}. \tag{1}$$

This leads to the solution of a linear system of equations which can be ill conditioned or possess several solutions. Therefore, regularized versions were introduced later. The linear least squares approach is optimal in the case of linear targets corrupted by Gaussian noise. Sometimes it is useful to find a linear function which does not minimize the least squares error, but, for example, the ℓ_1-error

$$\sum_{i=1}^{m} |\langle w, x_i\rangle + b - y_i| \;\to \min_{w,b}$$

which is more robust against outliers. This model with the constraint that the sum of the errors is equal to zero was already studied by Laplace in 1799; see Laplace

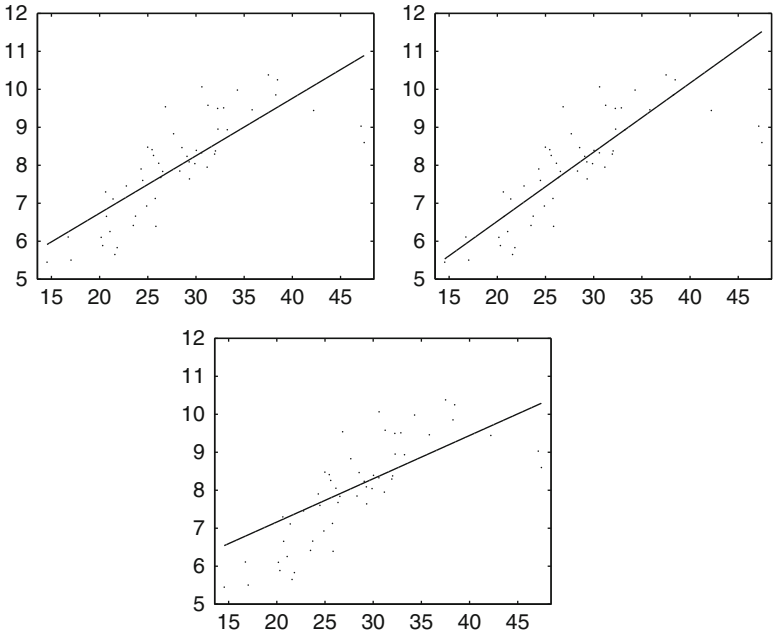


Fig. 2 Linear approximation with respect to the ℓ_2-, ℓ_1- and ℓ_∞-norm of the error (left to right). The ℓ_1 approximation is more robust against outliers while the ℓ_∞-norm takes them better into account

[61]. Another popular choice is the ℓ_∞-error

$$\max_{i=1,\ldots,m} |\langle w, x_i \rangle + b - y_i| \rightarrow \min_{w,b}$$

which better incorporates outliers. In contrast to the least squares method, the solutions of the ℓ_1- and ℓ_∞-problems cannot be computed via linear systems of equations but require to solve linear optimization problems. Figure 2 shows a one-dimensional example, where data are approximated by lines with minimal ℓ_2-, ℓ_1- and ℓ_∞ error norm, respectively. For more information on (regularized) least squares problems, the reader may consult, e.g., the books of Golub and Van Loan [45] and of Björck [10].

Regularized least squares methods which penalize the quadratic weight $\|w\|^2$ as in section "Linear Least Squares Classification and Regression" were examined under the name *ridge regression* by Hoerl and Kennard [49]. This method can be considered as a special case of the regularization theory for ill-posed problems developed by Tikhonov and Arsenin [102]. Others than the least squares loss function like the ϵ-*insensitive loss* were brought into play by Vapnik [105]. This loss function enforces a sparse representation of the weights in terms of so-called support vectors which are (small) subsets of the training samples $\{x_i : i = 1, \ldots, m\}$.

The simplest form of classification is *binary classification*, where one has just to separate between two classes. Linear hyperplanes $H(w, b)$ separating points, also called *linear discriminants* or *perceptrons*, were already studied by Fisher [41] and became interesting for neural network researchers in the early 1960s. One of the first algorithms that constructs a separating hyperplane for linearly separable points was Rosenblatt's perceptron; see Rosenblatt [84]. It is an iterative online and mistake-driven algorithm which starts with an initial weight guess w for the hyperplane and adapts the weight at each time a training point is misclassified by the current weight. If the data are linearly separable, the procedure converges, and the number of mistakes (number of updates of w) does not exceed $(2R/\gamma)^2$, where $R := \min_{i=1,\dots,m} \|x_i\|$ and γ is the smallest distance between a training point and the hyperplane. For linearly separable points, there exist various hyperplanes separating them.

An *optimal* hyperplane for linearly separable points in the sense that the minimal distance of the points from the plane becomes maximal was constructed as so-called generalized portrait algorithm by Vapnik and Lerner [108]. This learning method is also known as *linear hard margin support vector classifier*. The method was generalized to nonseparable points by Cortes and Vapnik [26] which leads to *soft margin classifiers*. Finally, the step from linear to nonlinear classifiers via feature maps was taken by Boser et al. [12]. Their idea to combine a linear algorithm with a kernel approach inspired the further examination of specific kernels for applications.

However, the theory of kernels and their applications is older than SVMs. Aronzajn [5], systematically developed the theory of RHKSs in the 1940s though it was discovered that many results were independently obtained by Povzner [83]. The work of Parzen [78] brought the RKHS to the fore in statistical problems; see also Kailath [54]. Kernels in pattern recognition were already applied by Aizerman et al. [1]. Empirical risk minimization over RKHSs was considered by Wahba [112] in connection with splines and by Poggio and Giriosi [81] in relation with neural networks. Schölkopf et al. [89] realized that the *kernel trick* works not only for SVMs but for many other methods as principal component analysis in unsupervised learning.

The invention of SVMs has led to a gigantic amount of developments in learning theory and practice. The size of this paper would be not enough to list the references on this topic. Beyond various applications, also advanced generalization results, suitable choices of kernels, efficient numerical methods in particular for large data sets, relations to other sparse representation methods, multi-class classification, and multitask learning were addressed. The reader will find some references in the corresponding sections.

3 Mathematical Modeling and Applications

Linear Learning

This section starts with linear classification and regression which provide the easiest algorithms to understand some of the main building blocks that appear also in the

more sophisticated nonlinear support vector machines. Moreover, concerning the classification task, this seems to be the best approach to explain its *geometrical background*. The simplest function to feed a classifier with or to use as an approximation of some unknown function in regression tasks is a linear (multivariate) function

$$f(x) = f_w(x) := \langle w, x \rangle, \quad x \in \mathcal{X} \subset \mathbb{R}^d. \tag{2}$$

Often it is combined with some appropriate real number b, i.e., one considers the linear polynomial $f(x) + b = \langle w, x \rangle + b$. In the context of learning, w is called *weight* vector and b *offset*, *intercept* or *bias*.

Linear Support Vector Classification

Let us consider binary classification first and postpone multi-class classification to section "Multi-class Classification and Multitask Learning." As binary classifier $F = F_{w,b} : \mathcal{X} \to \{-1, 1\}$, one can use

$$F(x) := \operatorname{sgn}(f_w(x) + b) = \operatorname{sgn}(\langle w, x \rangle + b)$$

with the agreement that $\operatorname{sgn}(0) := 0$. The hyperplane

$$H(w, b) := \{x : \langle w, x \rangle + b = 0\}$$

has the normal vector $w/\|w\|$, and the distance of a point $\tilde{x} \in \mathbb{R}^d$ to the hyperplane is given by

$$\left| \left\langle \frac{w}{\|w\|}, \tilde{x} \right\rangle + \frac{b}{\|w\|} \right|$$

see Fig. 3 left. In particular, $|b|/\|w\|$ is the distance of the hyperplane from the origin.

The training set Z consists of two classes labeled by ± 1 with indices $I_+ := \{i : y_i = 1\}$ and $I_- := \{i : y_i = -1\}$. The training set is said to be *separable by the hyperplane* $H(w, b)$ if $\langle w, x_i \rangle + b > 0$ for $i \in I_+$ and $\langle w, x_i \rangle + b < 0$ for $i \in I_-$, i.e.,

$$y_i(\langle w, x_i \rangle + b) > 0.$$

The points in Z are called (linearly) *separable* if there exists a hyperplane separating them. In this case, their distance from a separating hyperplane is given by

$$y_i \left(\left\langle \frac{w}{\|w\|}, x_i \right\rangle + \frac{b}{\|w\|} \right), \quad i = 1, \ldots, m.$$

The smallest distance of a point from the hyperplane

$$\gamma := \min_{i=1,\ldots,m} y_i \left(\left\langle \frac{w}{\|w\|}, x_i \right\rangle + \frac{b}{\|w\|} \right) \tag{3}$$

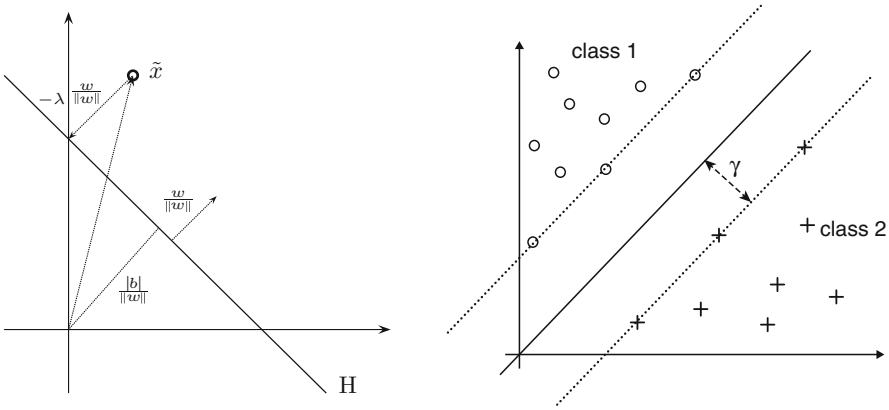


Fig. 3 *Left*: Hyperplane H with normal $w/\|w\|$ and distance $|b|/\|w\|$ from the origin. The distance of the point $\tilde{x}$ from the hyperplane is the value λ fulfilling $\langle w, \tilde{x} - \lambda w/\|w\| \rangle + b = 0$, i.e., $\lambda = (\langle w, \tilde{x} \rangle + b)/\|w\|$. *Right*: Linearly separable training set together with a separating hyperplane and the corresponding margin γ

is called *margin*. Figure 3, right, shows a separating hyperplane of two classes together with its margin. Of course for a separable training set, there may exist various separating hyperplanes. One way to ensure a unique solution is to pose additional requirements on the hyperplane in form of minimizing a cost functional.

Hard Margin Classifier One obvious way is to choose those separating hyperplane which has the maximal distance from the data, i.e., a maximal margin. The corresponding classifiers are called *maximal margin classifiers* or *hard margin classifiers*. The hyperplane and the corresponding half-spaces do not chance if the defining vectors is rescaled to $(c\, w, c\, b)$, $c > 0$. The so-called generalized portrait algorithm of Vapnik and Lerner [108], constructs a hyperplane that maximizes γ under the constraint $\|w\| = 1$. The same hyperplane can be obtained as follows: by (3), it holds

$$\gamma \, \|w\| = \min_{i=1,\dots,m} y_i \, (\langle w, x_i \rangle + b)$$

so that one can use the scaling

$$\gamma \, \|w\| = 1 \quad \Leftrightarrow \quad \gamma = \frac{1}{\|w\|}.$$

Now γ becomes maximal if and only if $\|w\|$ becomes minimal, and the scaling means that $y_i \, (\langle w, x_i \rangle + b) \geq 1$ for all $i = 1, \dots, m$. Therefore, the hard margin classifier aims to find parameters w and b solving the following quadratic optimization problem with linear constraints:

Linear SV hard margin classification (Primal problem)
$\frac{1}{2}\|w\|^2 \;\rightarrow\; \min_{w,b} \quad \text{subject to} \quad y_i\,(\langle w, x_i\rangle + b) \geq 1,\; i = 1,\ldots,m.$

If the training data are linearly separable, the problem has a unique solution. A brief introduction into quadratic programming methods is given in section "Quadratic Optimization." To transform the problem into its dual form, consider the Lagrangian

$$L(w,b,\alpha) = \frac{1}{2}\|w\|^2 + \sum_{i=1}^{m} \alpha_i\,(1 - y_i\,(\langle w, x_i\rangle + b))\,, \quad \alpha_i \geq 0.$$

Since

$$\frac{\partial L}{\partial w} = w - \sum_{i=1}^{m} \alpha_i\, y_i\, x_i \;=\; 0 \quad \Leftrightarrow \quad w \;=\; \sum_{i=1}^{m} \alpha_i\, y_i\, x_i\,, \tag{4}$$

$$\frac{\partial L}{\partial b} = \sum_{i=1}^{m} \alpha_i\, y_i \;=\; 0 \tag{5}$$

the Lagrangian can be rewritten as

$$L(w,b,\alpha) = -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} y_i\alpha_i\, y_j\alpha_j\, \langle x_i, x_j\rangle + \sum_{i=1}^{m} \alpha_i \tag{6}$$

and the dual problem becomes

Linear SV hard margin classification (Dual problem)
$\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} y_i\alpha_i\, y_j\alpha_j\, \langle x_i, x_j\rangle - \sum_{i=1}^{m} \alpha_i \rightarrow \min_{\alpha} \quad \text{subject to} \quad \sum_{i=1}^{m} y_i\alpha_i = 0,$ $\alpha_i \geq 0,\; i = 1,\ldots,m.$

Note that the dual maximization problem has been rewritten into a minimization problem by using that $\max \phi \;=\; \min -\phi$. Let 1_m denote the vector with m coefficients 1, $\alpha := (\alpha_i)_{i=1}^m$, $y := (y_i)_{i=1}^m$, $\mathbf{Y} := \mathrm{diag}(y_i)_{i=1}^m$ and

$$\mathbf{K} := \left(\langle x_i, x_j\rangle\right)_{i,j=1}^m. \tag{7}$$

Note that $\mathbf{K}$ is symmetric and positive semi-definite. The dual problem can be rewritten in *matrix-vector form* as

Linear SV hard margin classification (Dual problem)
$\frac{1}{2}\alpha^{\mathrm{T}}\mathbf{YKY}\alpha - \langle 1_m, \alpha\rangle \;\rightarrow\; \min_{\alpha} \quad \text{subject to} \quad \langle y, \alpha\rangle = 0,\; \alpha \geq 0.$

Let α^* be the minimizer of this dual problem. The intercept b does not appear in the dual problem, and one has to determine its optimal value in another way. By the Kuhn-Tucker conditions, the equations

$$\alpha_i^* \left(y_i (\langle w^*, x_i \rangle + b^*) - 1 \right) = 0, \quad i = 1, \ldots, m$$

hold true, so that $\alpha_i^* > 0$ is only possible for those training data with $y_i (\langle w^*, x_i \rangle + b^*) = 1$. These are exactly the (few) points having margin distance γ from the hyperplane $H(w^*, b^*)$. Define

$$I_S := \{i : \alpha_i^* > 0\}, \quad S := \{x_i : i \in I_S\}. \tag{8}$$

The vectors from S are called *support vectors*. In general $|S| \ll m$ and by (4) the optimal weight w^* and the optimal function f_{w^*} have a *sparse representation* in terms of the support vectors

$$w^* = \sum_{i \in I_S} \alpha_i^* y_i x_i, \quad f_{w^*}(x) = \sum_{i \in I_S} \alpha_i^* y_i \langle x_i, x \rangle. \tag{9}$$

Moreover,

$$b^* = y_i - \langle w^*, x_i \rangle = y_i - f_{w^*}(x_i), \quad i \in I_s \tag{10}$$

and hence, using (5),

$$\begin{aligned} \|w^*\|^2 &= \sum_{i \in I_S} \alpha_i^* y_i \sum_{j \in I_S} \alpha_j^* y_j \langle x_i, x_j \rangle = \sum_{i \in I_S} \alpha_i^* y_i f_{w^*}(x_i) \\ &= \sum_{i \in I_S} \alpha_i^* (1 - y_i b^*) = \sum_{i \in I_S} \alpha_i^* \end{aligned}$$

so that

$$\gamma = 1/\|w\| = \Big(\sum_{i \in I_S} \alpha_i^* \Big)^{-1/2}.$$

Soft Margin Classifier If the training data are not linearly separable which is the case in most applications, the hard margin classifier is not applicable. The extension of hard margin classifiers to the nonseparable case was done by Cortes and Vapnik [26] by bringing additional *slack variables* and a a parameter $C > 0$ into the constrained model:

Linear SV soft margin classification (Primal problem)			
$\frac{1}{2}\|w\|^2 + C \sum_{i=1}^{m} \xi_i$	$\to \min_{w,b,\xi}$	subject to	$y_i (\langle w, x_i \rangle + b) \geq 1 - \xi_i,$ $\xi_i \geq 0, \ i = 1, \ldots, m.$

For $C = \infty$, this is again the hard margin classifier model. As before, the *margin* is defined as $\gamma = 1/\|w^*\|$, where w^* is the solution of the above problem. If the slack variable fulfills $0 \leq \xi_i^* < 1$, then x_i is correctly classified, and in the case $y_i\,(\langle w^*, x_i\rangle + b^*) = 1 - \xi_i^*$ the distance of x_i from the hyperplane is $\gamma - \xi_i^*/\|w^*\|$. If $1 < \xi_i^*$, then one has a misclassification. By penalizing the sum of the slack variables, one tries to keep them small.

The above constrained minimization model can be rewritten as an unconstrained one by using a *margin-based loss function*. A function $L : \{-1, 1\} \times \mathbb{R} \to [0, \infty)$ is called *margin based* if there exists a representing function $l : \mathbb{R} \to [0, \infty)$ such that

$$L(y, t) = l(yt).$$

In soft margin classification the appropriate choice of a loss function is the *hinge loss function* l_h determined by

$$l_h(x) := \max\{0, 1 - x\}.$$

Then the *unconstrained* primal problem reads

Linear SV soft margin classification (Primal problem)

$$\tfrac{1}{2}\|w\|^2 + C \sum_{i=1}^{m} L_h\left(y_i, (\langle w, x_i\rangle + b)\right) \to \min_{w,b}.$$

The Lagrangian of the linear constrained problem has the form

$$L(w, b, \xi, \alpha) = \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{m} \xi_i + \sum_{i=1}^{m} \alpha_i \left(1 - \xi_i - y_i\left(\langle w, x_i\rangle + b\right)\right) - \sum_{i=1}^{m} \beta_i \xi_i,$$

where $\alpha_i, \beta_i \geq 0$. Partial differentiation of the Lagrangian with respect to w and b results in (4), (5) and with respect to ξ in

$$\frac{\partial L}{\partial \xi} = C\, 1_m - \alpha - \beta \;=\; 0.$$

Using these relations, the Lagrangian can be rewritten in the same form as in (6), and the dual problem becomes in matrix-vector form

Linear SV soft margin classification (Dual problem)

$$\tfrac{1}{2}\alpha^{\mathrm{T}}\mathbf{YKY}\alpha - \langle 1_m, \alpha\rangle \quad \text{subject to} \quad \langle y, \alpha\rangle = 0,\ 0 \leq \alpha \leq C.$$

Let α^* be the minimizer of the dual problem. Then the optimal weight w^* and f_{w^*} are again given by (9) and depend only on the few support vectors defined by (8). By the Kuhn-Tucker conditions, the equations

$$\alpha_i^* \left(y_i(\langle w^*, x_i\rangle + b^*) - 1 + \xi_i^*\right) = 0 \quad \text{and}$$

$$\beta_i^* \xi_i^* = (C - \alpha_i^*)\xi_i^* = 0, \quad i = 1, \ldots, m$$

hold true. For $0 < \alpha_i^* < C$, it follows that $\xi_i^* = 0$ and $y_i(\langle w^*, x_i \rangle + b^*) = 1$, i.e., the points x_i have margin distance $\gamma = 1/\|w^*\|$ from $H(w^*, b^*)$. Moreover,

$$b^* = y_i - \langle w^*, x_i \rangle, \quad i \in \tilde{I}_S := \{i : 0 < \alpha_i^* < C\}. \tag{11}$$

For $\alpha_i^* = C$, one concludes that $y_i(\langle w^*, x_i \rangle + b^*) = 1 - \xi_i^*$, i.e., x_i has distance $\gamma - \xi_i^*/\|w^*\|$ from the optimal hyperplane.

Linear Support Vector Regression

Of course one can also approximate unknown functions by linear (multivariate) polynomials of the form (2).

Hard Margin Regression The model for linear hard margin regression is given by

Linear SV hard margin regression (Primal problem)					
$\frac{1}{2}\|w\|^2 \to \min_{w,b}$	subject to	$\langle w, x_i \rangle + b - y_i$	$\leq$	ϵ,	
		$-\langle w, x_i \rangle - b + y_i$	$\leq$	ϵ,	$i = 1, \ldots, m$.

The constraints make sure that the test data y_i lie within an ϵ distance from the value $f(x_i) + b$ of the approximating linear polynomial. The Lagrangian reads

$$\begin{aligned} L(w, b, \xi^{\pm}, \alpha^{\pm}, \beta^{\pm}) = \frac{1}{2}\|w\|^2 &+ \sum_{i=1}^{m} \alpha_i^- \left(\langle w, x_i \rangle + b - y_i - \epsilon\right) \\ &+ \sum_{i=1}^{m} \alpha_i^+ \left(-\langle w, x_i \rangle - b + y_i - \epsilon\right), \end{aligned}$$

where $\alpha_i^{\pm} \geq 0$. Setting partial derivatives to zero leads to

$$\frac{\partial L}{\partial w} = w + \sum_{i=1}^{m} \left(\alpha_i^- - \alpha_i^+\right) x_i = 0 \quad \Leftrightarrow \quad w = \sum_{i=1}^{m} \left(\alpha_i^+ - \alpha_i^-\right) x_i, \tag{12}$$

$$\frac{\partial L}{\partial b} = \sum_{i=1}^{m} \left(\alpha_i^+ - \alpha_i^-\right) = 0. \tag{13}$$

Using these relations and setting

$$\alpha := \alpha^+ - \alpha^-,$$

the Lagrangian can be written as

$$L(w, b, \xi^{\pm}, \alpha^{\pm}, \beta^{\pm}) = -\frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j \langle x_i, x_j \rangle - \epsilon \sum_{i=1}^{m} \left(\alpha_i^+ + \alpha_i^-\right) + \sum_{i=1}^{m} y_i \alpha_i$$

and the dual problem becomes in *matrix-vector form*

Linear SV hard margin regression (Dual problem)
$\frac{1}{2}(\alpha^+ - \alpha^-)^{\mathrm{T}}\mathbf{K}(\alpha^+ - \alpha^-) + \epsilon\langle 1_m, \alpha^+ + \alpha^-\rangle - \langle y, \alpha^+ - \alpha^-\rangle \to \min_{\alpha^+,\alpha^-}$
subject to $\langle 1_m, \alpha^+ - \alpha^-\rangle = 0, \quad \alpha^{\pm} \geq 0.$

This is a quadratic optimization problem with linear constraints. Let $(\alpha^+)^*, (\alpha^-)^*$ be the solution of this problem and $\alpha^* = (\alpha^+)^* - (\alpha^-)^*$. Then, by (12), the optimal weight and the optimal function have in general a sparse representation in terms of the support vectors $x_i, i \in I_S$, namely,

$$w^* = \sum_{i \in I_S} \alpha_i^* x_i, \quad f_{w^*}(x) = \sum_{i \in I_{rS}} \alpha_i^* \langle x_i, x\rangle, \quad I_{rS} := \{i : \alpha_i^* \neq 0\}. \tag{14}$$

The corresponding Kuhn-Tucker conditions are

$$\begin{aligned} (\alpha_i^-)^* \left(\epsilon - \langle w^*, x_i\rangle - b^* + y_i\right) &= 0, \\ (\alpha_i^+)^* \left(\epsilon + \langle w^*, x_i\rangle + b^* - y_i\right) &= 0. \end{aligned} \tag{15}$$

If $(\alpha_i^-)^* > 0$ or $(\alpha_i^+)^* > 0$, then

$$b^* = y_i - \langle w^*, x_i\rangle + \epsilon, \quad b^* = y_i - \langle w^*, x_i\rangle - \epsilon,$$

respectively. Since both conditions cannot be fulfilled for the same index, it follows that $(\alpha_i^-)^*(\alpha_i^+)^* = 0$ and consequently, either $\alpha_i^* = (\alpha_i^+)^* \geq 0$ or $\alpha_i^* = -(\alpha_i^-)^* \leq 0$. Thus, one can obtain the intercept by

$$b^* = y_i - \langle w^*, x_i\rangle - \epsilon, \quad i \in I_S. \tag{16}$$

Soft Margin Regression Relaxing the constraints in the hard margin model leads to the following linear soft margin regression problem with $C > 0$:

Linear SV soft margin regression (Primal problem)		
$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^m \left(\xi_i^+ + \xi_i^-\right) \to \min_{w,b,\xi_i^{\pm}}$	subject to $\langle w, x_i\rangle + b - y_i$	$\leq \epsilon + \xi_i^-,$
	$-\langle w, x_i\rangle - b + y_i$	$\leq \epsilon + \xi_i^+,$
	ξ_i^+, ξ_i^-	$\geq 0.$

For $C = \infty$, this recovers the linear hard margin regression problem. The above constrained model can be rewritten as an unconstrained one by using a *distance-based loss function*. A function $L : \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ is called *distance based* if there exists a representing function $l : \mathbb{R} \to [0, \infty)$ such that

$$L(y, t) = l(y - t).$$

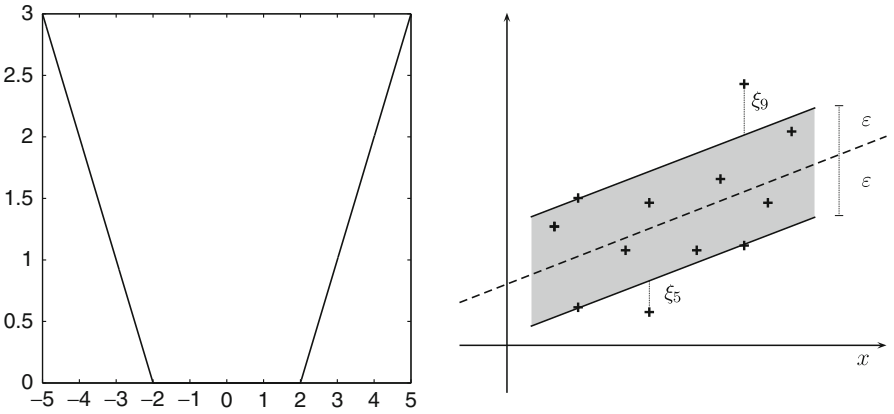


Fig. 4 Vapnik's ϵ-insensitive loss function for $\epsilon = 2$ (*left*). Example of linear SV soft margin regression (*right*)

In soft margin regression, the appropriate choice is Vapnik's *ϵ-insensitive loss function* defined by

$$l_\epsilon(x) := \max\{0, |x| - \epsilon\}.$$

The function l_ϵ is depicted in Fig. 4, left. Then the *unconstrained* primal model reads

Linear SV soft margin regression (Primal problem)

$$\tfrac{1}{2}\|w\|^2 + C \sum_{i=1}^{m} L_\epsilon(y_i, \langle w, x_i\rangle + b) \;\to\; \min_{w,b}.$$

The Lagrangian of the constrained problem is given by

$$\begin{aligned} L(w, b, \xi^\pm, \alpha^\pm, \beta^\pm) = \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{m} (\xi_i^+ + \xi_i^-) \\ + \sum_{i=1}^{m} \alpha_i^- \left(\langle w, x_i\rangle + b - y_i - \epsilon - \xi_i^-\right) \\ + \sum_{i=1}^{m} \alpha_i^+ \left(-\langle w, x_i\rangle - b + y_i - \epsilon - \xi_i^+\right) \\ - \sum_{i=1}^{m} \beta_i^+ \xi_i^+ - \sum_{i=1}^{m} \beta_i^- \xi_i^-, \end{aligned}$$

where $\alpha_i^\pm \geq 0$ and $\beta_i^\pm \geq 0$, $i = 1, \ldots, m$. Setting the partial derivatives to zero leads to (12), (13) and

$$\frac{\partial L}{\partial \xi^+} = C\, 1_m - \alpha^+ - \beta^+ = 0, \quad \frac{\partial L}{\partial \xi^-} = C\, 1_m - \alpha^- - \beta^- = 0.$$

Using these relations the Lagrangian can be written exactly as in the hard margin problem, and the dual problem becomes in *matrix-vector form*

Linear SV soft margin regression (Dual problem)

$$\tfrac{1}{2}(\alpha^+ - \alpha^-)^{\mathrm{T}}\mathbf{K}(\alpha^+ - \alpha^-) + \epsilon\langle 1_m, \alpha^+ + \alpha^-\rangle - \langle y, \alpha^+ - \alpha^-\rangle \to \min_{\alpha^+,\alpha^-}$$

$$\text{subject to} \quad \langle 1_m, \alpha^+ - \alpha^-\rangle = 0, \quad 0 \le \alpha^+, \alpha^- \le C$$

If $(\alpha^+)^*, (\alpha^-)^*$ are the solution of this problem and $\alpha^* = (\alpha^+)^* - (\alpha^-)^*$, then the optimal weight w^* and the optimal function f_{w^*} are given by (14). The corresponding Kuhn-Tucker conditions are

$$\begin{aligned}
(\alpha_i^-)^* \left(\epsilon + (\xi_i^-)^* - \langle w^*, x_i\rangle - b^* + y_i\right) &= 0,\\
(\alpha_i^+)^* \left(\epsilon + (\xi_i^+)^* + \langle w^*, x_i\rangle + b^* - y_i\right) &= 0,\\
(C - (\alpha_i^+)^*)(\xi_i^+)^* = 0, \quad (C - (\alpha_i^-)^*)(\xi_i^-)^* &= 0, \quad i = 1, \ldots, m.
\end{aligned}$$

If $0 < (\alpha_i^+)^*$ or $0 < (\alpha_i^-)^*$, then

$$b^* = y_i - \langle w^*, x_i\rangle + \epsilon + \xi_i^+, \quad b^* = y_i - \langle w^*, x_i\rangle - \epsilon - \xi_i^-,$$

respectively. Both equations cannot be fulfilled at the same time so that one can conclude that either $\alpha_i^* = (\alpha_i^+)^* \ge 0$ or $\alpha_i^* = -(\alpha_i^-)^* \le 0$. In case $\alpha_i^* = (\alpha_i^+)^* < C$, this results in the intercept

$$b^* = y_i - \langle w^*, x_i\rangle - \epsilon, \quad i \in \tilde{I}_S. \tag{17}$$

Linear Least Squares Classification and Regression

Instead of the hinge loss function for classification and the ϵ-insensitive loss function for regression, other loss functions can be used. Popular margin-based and distance-based loss functions are the *logistic loss* for classification and regression

$$l(yt) := \ln(1 + e^{-yt}) \quad \text{and} \quad l(y - t) := -\ln \frac{4e^{y-t}}{(1 + e^{y-t})^2},$$

respectively. In contrast to the loss functions in the previous subsections, logistic loss functions are differentiable in t so that often standard methods as gradient descent methods or Newton (like) methods can be applied for computing the minimizers of the corresponding problems. For details, see, e.g., the book of Mitchell [73] or of Hastie et al. [47]. Other loss functions for regression are the *pinball loss*, the *Huber function*, and the *p-th power absolute distance loss* $|y - t|^p$, $p > 0$. For $p = 2$, the latter is the *least squares loss*

$$l_{\mathrm{lsq}}(y-t) = (y-t)^2.$$

Since $(y-t)^2 = (1-yt)^2$ for $y \in \{-1,1\}$, the least squares loss is also margin based, and one can handle least squares classification and regression using just the same model with $y \in \{-1,1\}$ for classification and $y \in \mathbb{R}$ for regression. In the *unconstrained* form, one has to minimize

Linear LS classification/regression (Primal problem)

$$\tfrac{1}{2}\|w\|^2 + \tfrac{C}{2}\sum_{i=1}^{m} \underbrace{(\langle w, x_i\rangle + b - y_i)^2}_{L_{\mathrm{lsq}}(y_i,\langle w,x_i\rangle + b)} \;\to\; \min_{w,b}.$$

This model was published as *ridge regression* by Hoerl and Kennard [49] and is a regularized version of the Gaussian model (1). Therefore, it can be considered as a special case of regularization theory introduced by Tikhonov and Arsenin [102]. The minimizer can be computed via a linear system of equations. To this end, rewrite the unconstrained problem in matrix-vector form

$$\frac{1}{2}\|w\|^2 + \frac{C}{2}\|\mathbf{X}^{\mathrm{T}} w + b1_m - y\|^2 \;\to\; \min_{w,b},$$

where

$$\mathbf{X} := (x_1 \;\ldots\; x_m) = \begin{pmatrix} x_{1,1} & \ldots & x_{m,1} \\ & \vdots & \\ x_{1,d} & \ldots & x_{m,d} \end{pmatrix}.$$

Setting the gradient (with respect to w and b) to zero, one obtains

$$\begin{aligned} 0 &= \frac{1}{C} w + \mathbf{X}\mathbf{X}^{\mathrm{T}} w + b\mathbf{X}1_m - \mathbf{X}y, \\ 0 &= 1_m^{\mathrm{T}}\mathbf{X}^{\mathrm{T}} w - 1_m^{\mathrm{T}} y + mb \quad \Leftrightarrow \quad b = \bar{y} - \langle w, \bar{x}\rangle, \end{aligned} \tag{18}$$

where $\bar{y} := \frac{1}{m}\sum_{i=1}^m y_i$ and $\bar{x} := \frac{1}{m}\sum_{i=1}^m x_i$. Hence b and w can be obtained by solving the linear system of equations

$$\left(\begin{array}{c|c} 1 & \bar{x}^{\mathrm{T}} \\ \hline \bar{x} & \frac{1}{m}\mathbf{X}\mathbf{X}^{\mathrm{T}} + \frac{1}{mC} I \end{array}\right) \begin{pmatrix} b \\ w \end{pmatrix} = \begin{pmatrix} \bar{y} \\ \frac{1}{m}\mathbf{X}y \end{pmatrix}. \tag{19}$$

Instead of the above problem, one solves in general the "centered" problem

Linear LS classification/regression *in centered version* (Primal problem)

$$\tfrac{1}{2}\|\tilde{w}\|^2 + \tfrac{C}{2}\sum_{i=1}^{m} (\langle \tilde{w}, \tilde{x}_i\rangle + \tilde{b} - y_i)^2 \;\to\; \min_{\tilde{w},\tilde{b}},$$

where $\tilde{x}_i := x_i - \bar{x}$, $i = 1, \ldots, m$. The advantage is that $\bar{\tilde{x}} = 0_m$, where 0_m is the vector consisting of m zeros. Thus, (19) with $\tilde{x}_i$ instead of x_i becomes a separable system, and one obtains immediately that $\tilde{b}^* = \bar{y}$ and that $\tilde{w}^*$ follows by solving the linear system with positive definite, symmetric coefficient matrix

$$\left(\tilde{\mathbf{X}}\tilde{\mathbf{X}}^{\mathrm{T}} + \frac{1}{C} I\right) w = \tilde{\mathbf{X}} y.$$

This means that $\tilde{w}$ is just the solution of the centered primal problem without intercept. Finally, one can check by the following argument that indeed $w^* = \tilde{w}^*$:

$$\begin{aligned}
(\tilde{w}^*, \tilde{b}^*) &:= \operatorname{argmin}_{\tilde{w},\tilde{b}} \left\{ \frac{1}{2}\|\tilde{w}\|^2 + \frac{C}{2} \sum_{i=1}^{m} (\langle \tilde{w}, \tilde{x}_i \rangle + \tilde{b} - y_i)^2 \right\}, \\
\tilde{w}^* &= \operatorname{argmin}_{\tilde{w}} \left\{ \frac{1}{2}\|\tilde{w}\|^2 + \frac{C}{2} \sum_{i=1}^{m} (\langle \tilde{w}, \tilde{x}_i \rangle + \bar{y} - y_i)^2 \right\} \\
&= \operatorname{argmin}_{\tilde{w}} \left\{ \frac{1}{2}\|\tilde{w}\|^2 + \frac{C}{2} \sum_{i=1}^{m} (\langle \tilde{w}, x_i \rangle + \bar{y} - \langle \tilde{w}, \bar{x} \rangle - y_i)^2 \right\}
\end{aligned}$$

and with (18) on the other hand

$$\begin{aligned}
(w^*, b^*) &:= \operatorname{argmin}_{w,b} \left\{ \frac{1}{2}\|w\|^2 + \frac{C}{2} \sum_{i=1}^{m} (\langle w, x_i \rangle + b - y_i)^2 \right\}, \\
w^* &= \operatorname{argmin}_{w} \left\{ \frac{1}{2}\|w\|^2 + \frac{C}{2} \sum_{i=1}^{m} (\langle w, x_i \rangle + \bar{y} - \langle w, \bar{x} \rangle - y_i)^2 \right\}.
\end{aligned}$$

Note that $\mathbf{X}^{\mathrm{T}}\mathbf{X} = \mathbf{K} \in \mathbb{R}^{m,m}$, but this is not the coefficient matrix in (19). When turning to nonlinear methods in section "Nonlinear Learning," it will be essential to work with $\mathbf{K} = \mathbf{X}^{\mathrm{T}}\mathbf{X}$ instead of $\mathbf{X}\mathbf{X}^{\mathrm{T}}$. This can be achieved by switching to the dual setting. In the following, this dual approach is shown although it makes often not sense *for the linear setting* since the size of the matrix $\mathbf{K}$ is in general larger than those of $\mathbf{X}\mathbf{X}^{\mathrm{T}} \in \mathbb{R}^{d,d}$. First, one reformulates the primal problem into a constrained one:

Linear LS classification/regression (Primal problem)
$\frac{1}{2}\|w\|^2 + \frac{C}{2} \sum_{i=1}^{m} \xi_i^2 \to \min_{w,b,\xi}$ subject to $\langle w, x_i \rangle + b - y_i = \xi_i,\ i = 1, \ldots, m.$

The Lagrangian reads

$$L(w, b, \xi, \alpha) = \frac{1}{2}\|w\|^2 + \frac{C}{2} \sum_{i=1}^{m} \xi_i^2 - \sum_{i=1}^{m} \alpha_i \left(\langle w, x_i \rangle + b - y_i - \xi_i\right)$$

and

$$\frac{\partial L}{\partial w} = w - \sum_{i=1}^{m} \alpha_i x_i \;=\; 0 \quad \Leftrightarrow \quad w \;=\; \sum_{i=1}^{m} \alpha_i x_i\,,$$

$$\frac{\partial L}{\partial b} = \sum_{i=1}^{m} \alpha_i \;=\; 0,$$

$$\frac{\partial L}{\partial \xi} = C\,\xi + \alpha \;=\; 0.$$

The equality constraint in the primal problem together with the above equalities leads to the following linear system of equations to determine the optimal α^* and b^*:

$$\left(\begin{array}{c|c} 0 & 1_m^{\mathrm{T}} \\ \hline 1_m & \mathbf{K} + \frac{1}{C} I \end{array}\right) \begin{pmatrix} b \\ \alpha \end{pmatrix} = \begin{pmatrix} 0 \\ y \end{pmatrix}. \tag{20}$$

The optimal weight and the corresponding optimal function read

$$w^* = \sum_{i=1}^{m} \alpha_i^* x_i\,, \qquad f_{w^*}(x) = \sum_{i=1}^{m} \alpha_i^* \langle x_i\,, x \rangle. \tag{21}$$

In general there is no sparse representation with respect to the vectors x_i; see also Theorem 8.36 in the book of Steinwart and Christman [93]. Therefore, this method is not called a support vector method in this paper. Finally, note that the centered approach also helps to avoid the intercept in the dual approach. Since this is no longer true when turning to the nonlinear setting, the intercept is kept here.

Nonlinear Learning

A linear form of a decision or regression function may not be suitable for the task at hand. Figure 5 shows two examples, where a linear classifier is not appropriate.

A basic idea to handle such problems was proposed by Boser et al. [12] and consists of the following two steps which will be further explained in the rest of this subsection:

1. Mapping of the input data $X \subset \mathcal{X}$ into a *feature space* $\Phi(\mathcal{X}) \subset \ell_2(I)$, where I is a countable (possibly finite) index set, by a nonlinear *feature map*

$$\Phi : \mathcal{X} \to \ell_2(I).$$

2. Application of the linear classification/regression model to the feature set

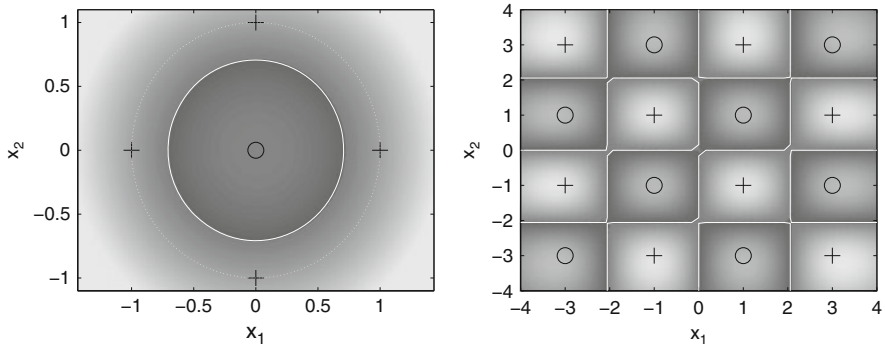


Fig. 5 Two sets, where linear classification is not appropriate

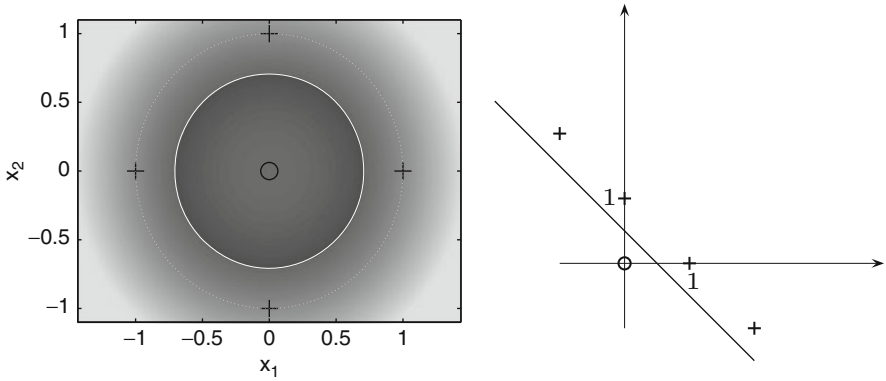


Fig. 6 Linearly nonseparable training data in the original space $\mathcal{X} \subset \mathbb{R}^2$ (*left*) and become separable in the feature space $\Phi(\mathcal{X})$, where $\Phi(x_1, x_2) := (x_1^2, x_2^2)$ (*right*)

$$\{(\Phi(x_1), y_1), \ldots, (\Phi(x_m), y_m)\}.$$

This means that instead of a linear function (2), we are searching for a function of the form

$$f(x) = f_w(x) := \langle w, \Phi(x) \rangle_{\ell_2(I)} \tag{22}$$

now. This nonlinear function on $\mathcal{X}$ becomes linear in the feature space $\Phi(\mathcal{X})$.

Figure 6 shows an example of a feature map. In this example, the set $\{(x_i, y_i) : i = 1, \ldots, 5\}$ is not linearly separable while $\{(\Phi(x_i), y_i) : i = 1, \ldots, 5\}$ is linearly separable. In practical applications, in contrast to this example, the feature map often maps into a higher dimensional, possibly also infinite dimensional space.

Together with the so-called kernel trick to avoid the direct work with the feature map Φ, this approach results in the successful *support vector machine* (SVM).

Kernel Trick

In general, one avoids to work directly with the feature map by dealing with the dual problem and applying the so-called kernel trick. For a feature map Φ, define a *kernel* $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ associated with Φ by

$$K(x,t) := \langle \Phi(x), \Phi(t) \rangle_{l_2(I)}. \tag{23}$$

More precisely, in practice one often follows the opposite way, namely, one starts with a suitable kernel which is known to have a representation of the form (23) without knowing Φ explicitly.

A frequently applied group of kernels are continuous, symmetric, positive (semi-) definite kernels like the Gaussian

$$K(x,t) = e^{-\|x-t\|^2/c^2}.$$

These kernels, which are also called *Mercer kernels*, will be considered in detail in section "Reproducing Kernel Hilbert Spaces." By Mercer's theorem, it can be shown that a Mercer kernel possesses a representation

$$K(x,t) = \sum_{j \in I} \sqrt{\lambda_j}\psi_j(x)\sqrt{\lambda_j}\psi_j(t), \quad x,t \in \mathcal{X}$$

with L_2-orthonormal functions ψ_j and positive values λ_j, where the right-hand side converges uniformly. Note that the existence of such a representation is clear, but in general without knowing the functions ψ_j explicitly. The set $\{\varphi_j := \sqrt{\lambda_j}\psi_j : j \in I\}$ forms an orthonormal basis of a *reproducing kernel Hilbert space* (RKHS) $\mathcal{H}_K$. These spaces are considered in more detail in section "Reproducing Kernel Hilbert Spaces." Then the feature map is defined by

$$\Phi(x) := \big(\varphi_j(x)\big)_{j \in I} = \left(\sqrt{\lambda_j}\psi_j(x)\right)_{j \in I}.$$

Using the orthonormal basis, one knows that for any $f \in \mathcal{H}_K$ there exists a unique sequence $w = w_f := (w_j)_{j \in I} \in \ell_2(I)$ such that

$$f(x) = \sum_{j \in I} w_j \varphi_j(x) = \langle w, \Phi(x) \rangle, \quad \text{and} \quad w_j = \langle f, \varphi_j \rangle_{\mathcal{H}_K}, \tag{24}$$

where the convergence is uniformly. Conversely every sequence $w \in \ell_2(I)$ defines a function f_w lying in $\mathcal{H}_K$ by (24). Moreover, Parseval's equality says that

$$\|f_w\|_{\mathcal{H}_K} := \|w\|_{\ell_2(I)}. \tag{25}$$

For nonlinear classification and regression purposes, one can follow exactly the lines of section "Linear Learning" except that one has to work in $\Phi(\mathcal{X})$ instead of $\mathcal{X}$. Using (22) instead of (2) and

$$\mathbf{K} := \left(\langle \Phi(x_i), \Phi(x_j)\rangle_{\ell_2(I)}\right)_{i,j=1}^m = \left(K(x_i, x_j)\right)_{i,j=1}^m \tag{26}$$

instead of the kernel matrix $\mathbf{K} := \left(\langle x_i, x_j\rangle\right)_{i,j=1}^m$ in (7), the linear models from section "Linear Learning" can be rewritten as in the following subsections. Note again that $\mathbf{K}$ is positive semi-definite.

Support Vector Classification

In the following, the linear classifiers are generalized to feature spaces.

Hard Margin Classifier The hard margin classification model is

SVM hard margin classification (Primal problem)

$$\tfrac{1}{2}\|w\|_{\ell_2(I)}^2 \;\to\; \min_{w,b} \quad \text{subject to} \quad y_i\left(\langle w, \Phi(x_i)\rangle_{\ell_2(I)} + b\right) \geq 1,\ i = 1, \ldots, m.$$

Interestingly, if Φ is associated with a Mercer kernel K, then $f(x) = \langle w, \Phi(x)\rangle_{\ell_2(I)}$ lies in the RKHS $\mathcal{H}_K$, and the model can be rewritten using (25) from the *point of view of RKHS* as

SVM hard margin classification *in RKHS* (Primal problem)

$$\tfrac{1}{2}\|f\|_{\mathcal{H}_K}^2 \;\to\; \min_{f \in \mathcal{H}_K} \quad \text{subject to} \quad y_i\left(f(x_i) + b\right) \geq 1,\ i = 1, \ldots, m.$$

The dual problem reads

SVM hard margin classification (Dual problem)

$$\tfrac{1}{2}\sum_{i=1}^m \sum_{j=1}^m y_i \alpha_i y_j \alpha_j \langle \Phi(x_i), \Phi(x_j)\rangle_{\ell_2(I)} - \sum_{i=1}^m \alpha_i \;\to\; \min_{\alpha} \text{ subject to } \sum_{i=1}^m y_i \alpha_i = 0,$$
$$\alpha_i \geq 0,\ i = 1, \ldots, m.$$

and with $\mathbf{K}$ defined by (26) the *matrix-vector form of the dual problem* looks as those for the linear hard margin classifier.

Let α^* be the minimizer of the dual problem. Then, by (9) together with the feature space modification, the optimal weight and the function f_{w^*} become

$$w^* = \sum_{i \in I_S} \alpha_i^* y_i \Phi(x_i), \quad f_{w^*}(x) = \sum_{i \in I_S} \alpha_i^* y_i \langle \Phi(x_i), \Phi(x)\rangle_{\ell_2(I)} = \sum_{i \in I_S} \alpha_i^* y_i K(x_i, x). \tag{27}$$

Thus, one can compute f_{w^*} knowing only the kernel and not the feature map itself. One property of a Mercer kernel used for learning purposes should be that it can be simply evaluated at points from $\mathcal{X} \times \mathcal{X}$. This is, for example, the case for the Gaussian. Finally, using (10) in the feature space, the intercept can be computed by

$$b^* = y_i - \langle w^*, \Phi(x_i)\rangle_{\ell_2(I)} = y_i - \sum_{j\in I_S} \alpha_j^* y_j K(x_j, x_i), \quad i \in I_S$$

and the margin $\gamma = 1/\|w^*\|^2_{\ell_2(I)}$ by using $\|w^*\|^2_{\ell_2(I)} = (\alpha^*)^{\mathrm{T}}\mathbf{YKY}\alpha^*$.

Soft Margin Classifier The soft margin classification model in the feature space is

SVM soft margin classification (Primal problem)

$$\tfrac{1}{2}\|w\|^2_{\ell_2(I)} + C\sum_{i=1}^m \xi_i \to \min_{w,b,\xi} \text{ subject to } y_i\left(\langle w, \Phi(x_i)\rangle_{\ell_2(I)} + b\right) \geq 1 - \xi_i, \quad i = 1,\dots,m$$
$$\xi_i \geq 0,\ i = 1,\dots,m.$$

If Φ is associated with a Mercer kernel K, the corresponding unconstrained version reads in the RKHS

SVM soft margin classification *in RKHS* (Primal problem)

$$\tfrac{1}{2}\|f\|^2_{\mathcal{H}_K} + C\sum_{i=1}^m L_h\left(y_i, f(x_i) + b\right)_+ \to \min_{f\in\mathcal{H}_K}.$$

With $\mathbf{K}$ defined by (26), the *matrix-vector form of the dual problem* looks as those for the linear soft margin classifier. The function f_{w^*} reads as in (27), and, using (11), the intercept can be computed by

$$b^* = y_i - \langle w^*, \Phi(x_i)\rangle_{\ell_2(I)} = y_i - \sum_{j\in\tilde{I}_S} \alpha_j^* y_j K(x_j, x_i), \quad i \in \tilde{I}_S.$$

Support Vector Regression

In the following, the linear regression models are generalized to feature spaces.

Hard Margin Regression One obtains

SVM hard margin regression (Primal problem)

$$\tfrac{1}{2}\|w\|^2_{\ell_2(I)} \to \min_{w,b} \text{ subject to } \langle w, \Phi(x_i)\rangle_{\ell_2(I)} + b - y_i \leq \epsilon,$$
$$-\langle w, \Phi(x_i)\rangle_{\ell_2(I)} - b + y_i \leq \epsilon, \quad i = 1,\dots,m.$$

If Φ is associated with a Mercer kernel K, then $f(x) = \langle w, \Phi(x)\rangle_{\ell_2(I)}$ lies in the RKHS $\mathcal{H}_K$, and the model can be rewritten using (25) from the *point of view of RKHS* as

SVM hard margin regression *in RKHS* (Primal problem)

$$\tfrac{1}{2}\|f\|^2_{\mathcal{H}_K} \to \min_{f\in\mathcal{H}_K} \text{ subject to } f(x_i) + b - y_i \leq \epsilon,$$
$$-f(x_i) - b + y_i \leq \epsilon, \quad i = 1,\dots,m.$$

The dual problem reads in matrix-vector form as the dual problem for the linear SV hard margin regression except that we have to use the kernel matrix $\mathbf{K}$

defined by (26). Let $(\alpha^+)^*, (\alpha^-)^*$ be the solution of this dual problem and $\alpha^* = (\alpha^+)^* - (\alpha^-)^*$. Then one can compute the optimal function f_{w^*} using (14) with the corresponding feature space modification as

$$f_{w^*}(x) = \sum_{i \in I_{rS}} \alpha_i^* K(x_i, x). \tag{28}$$

One obtains sparse representations in terms of the support vectors. By (16), the intercept can be computed by

$$b^* = y_i - \sum_{j \in I_{rS}} \alpha_j^* K(x_j, x_i) - \epsilon, \quad i \in I_S.$$

Soft Margin Regression In the feature space, the soft margin regression model is

SVM soft margin regression (Primal problem)

$$\tfrac{1}{2}\|w\|^2_{\ell_2(I)} + C \sum_{i=1}^m (\xi_i^+ + \xi_i^-) \;\to\; \min_{w,b,\xi_i^\pm} \text{ s.t. } \begin{aligned} \langle w, \Phi(x_i)\rangle_{\ell_2(I)} + b - y_i &\le \epsilon + \xi_i^-, \\ -\langle w, \Phi(x_i)\rangle_{\ell_2(I)} - b + y_i &\le \epsilon + \xi_i^+, \\ \xi_i^+, \xi_i^- &\ge 0. \end{aligned}$$

Having a feature map associated with a Mercer kernel K, the corresponding unconstrained problem can be written as the following minimization problem in the RKHS $\mathcal{H}_K$:

SVM soft margin regression *in RKHS* (Primal problem)

$$\tfrac{1}{2}\|f\|_{\mathcal{H}_K} + C \sum_{i=1}^m L_\epsilon(y_i, f(x_i) + b) \;\to\; \min_{f \in \mathcal{H}_K}.$$

The dual problem looks as the dual problem for linear SV soft margin regression but with kernel (26). From the solution of the dual problem $(\alpha^+)^*, (\alpha^-)^*$, one can compute the optimal function f_{w^*} as in (28) and the optimal intercept using (17) as

$$b^* = y_i - \sum_{j \in I_{rS}} \alpha_j^* K(x_j, x_i) - \epsilon, \quad i \in \tilde{I}_S.$$

Figure 7, left, shows an SVM soft margin regression function for the data in Fig. 2.

Relations to Sparse Approximation in RKHSs, Interpolation by Radial Basis Functions, and Kriging

SVM regression is related to various tasks in approximation theory. Some of them will be sketched in the following.

Relation to Sparse Approximation in RKHSs Let $\mathcal{H}_K$ be a RKHS with kernel K. Consider the problem of finding for an unknown function $\tilde{f} \in \mathcal{H}_K$ with given $\tilde{f}(x_i) = y_i, i = 1, \ldots, m$ an approximating function of the form

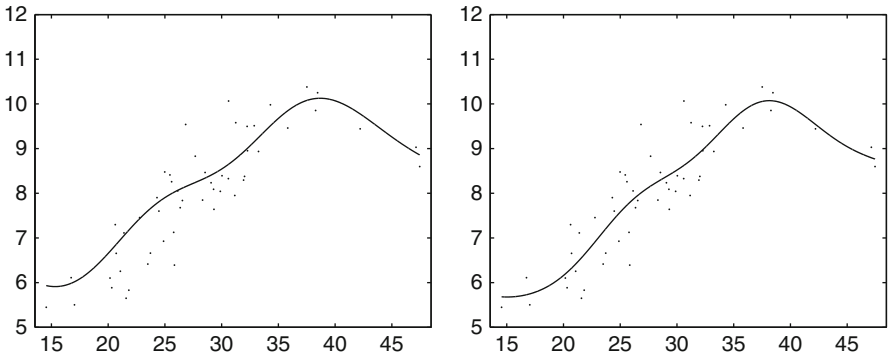


Fig. 7 SVM regression using the Gaussian with $c = 8$ for the data in Fig. 2. SVM soft margin regression curve with $C = 0.5$ and $\epsilon = 0.2$ (*left*). Least squares SVM regression curve with $C = 40$

$$f(x) := \sum_{i=1}^{m} \alpha_i K(x, x_i) \in \mathcal{H}_K \tag{29}$$

with only few summands. A starting point would be to minimize the $\mathcal{H}_K$-norm of the error and to penalize the so-called ℓ_0-norm of α given by $\|\alpha\|_0 := |\{i : \alpha_i \neq 0\}|$ to enforce sparsity. Unfortunately, the complexity when solving such ℓ_0-penalized problems increases exponentially with m. One remedy is to replace the ℓ_0-norm by the ℓ_1-norm, i.e., to deal with

$$\frac{1}{2}\|\tilde{f}(x) - \sum_{i=1}^{m} \alpha_i K(x, x_i)\|_{\mathcal{H}_K}^2 + \epsilon\|\alpha\|_1 \;\to\; \min_{\alpha}, \tag{30}$$

where $\epsilon > 0$. This problem and its relation to support vector regression were considered by Girosi [44] and Evgeniou et al. [37]. Using the relations in RKHS from section "Reproducing Kernel Hilbert Spaces," in particular the reproducing property (H2), this problem becomes

$$\frac{1}{2}\alpha^{\mathrm{T}}\mathbf{K}\alpha - \sum_{i=1}^{m} \alpha_i y_i + \frac{1}{2}\|\tilde{f}\|_{\mathcal{H}_K}^2 + \epsilon\|\alpha\|_1 \;\to\; \min_{\alpha}, \tag{31}$$

where $\mathbf{K}$ is defined by (26). With the splitting

$$\alpha_i = \alpha_i^+ - \alpha_i^-,\ \alpha_i^{\pm} \geq 0,\ \alpha_i^+ \alpha_i^- = 0,\ i = 1, \ldots, m$$

and consequently $|\alpha_i| = \alpha_i^+ + \alpha_i^-$, the sparse approximation model (30) has finally the form of the *dual problem of the SVM hard margin regression without intercept*:

SVM hard margin regression *without intercept* (Dual problem)

$$\tfrac{1}{2}(\alpha^+ - \alpha^-)^{\mathrm{T}}\mathbf{K}(\alpha^+ - \alpha^-) + \epsilon\langle 1_m, \alpha^+ + \alpha^-\rangle - \langle y, \alpha^+ - \alpha^-\rangle \;\to \min_{\alpha^+,\alpha^-}$$

$$\text{subject to} \quad \alpha^{\pm} \geq 0.$$

Note that for $\epsilon > 0$ the additional constraints $\alpha_i^+\alpha_i^- = 0$, $i = 1,\ldots,m$ are automatically fulfilled by the minimizer since otherwise, the Kuhn-Tucker conditions (15) without intercept would imply the contradiction $f(x_i) = y_i + \epsilon = y_i - \epsilon$.

Relation to the Interpolation by Radial Basis Functions For $\epsilon = 0$, problem (30), resp. (31) becomes

$$F(\alpha) := \frac{1}{2}\alpha^{\mathrm{T}}\mathbf{K}\alpha - \alpha^{\mathrm{T}}y \;\to \min_{\alpha}.$$

If $\mathbf{K}$ is positive definite, the solution of this problem is given by the solution of

$$\nabla F(\alpha) = \mathbf{K}\alpha - y = 0, \quad \Leftrightarrow \quad \mathbf{K}\alpha = y$$

and the approximating function f reads

$$f(x) = \langle \mathbf{K}^{-1}y, (K(x, x_i))_{i=1}^m \rangle. \tag{32}$$

This is just the solution of the *interpolation problem* to find f of the form (29) such that $f(x_i) = y_i$ for all $i = 1,\ldots,m$. If the kernel K of the positive definite matrix arises from a radial basis function $\kappa(x) = k(\|x - t\|^2)$, i.e., $K(x,t) = \kappa(x - t)$ as, e.g., from a Gaussian or an inverse multiquadric described in section "Reproducing Kernel Hilbert Spaces," this interpolation problem is called *interpolation by radial basis function*.

If the kernel K arises from a conditionally positive definite radial function κ of order ν, e.g., from a thin plate spline, the matrix $\mathbf{K}$ is in general not positive semi-definite. In this case, it is useful to replace the function f in (29) by

$$f(x) := \sum_{i=1}^{m} \alpha_i K(x, x_i) + \sum_{k=1}^{n} \beta_k p_k(x),$$

where n is the dimension of the polynomial space $\Pi_{\nu-1}(\mathbb{R}^d)$ and $\{p_k : k = 1,\ldots,n\}$ is a basis of $\Pi_{\nu-1}(\mathbb{R}^d)$. The additional degrees of freedom in the interpolation problem $f(x_i) = y_i$, $i = 1,\ldots,m$ are compensated by adding the new conditions

$$\sum_{i=1}^{m} \alpha_i p_k(x_i) = 0, \quad k = 1,\ldots,n.$$

This leads to the final problem of finding $\alpha := (\alpha_i)_{i=1}^m$ and $\beta := (\beta_k)_{k=1}^n$ such that

$$\begin{pmatrix} \mathbf{K} & \mathbf{P} \\ \mathbf{P}^\mathrm{T} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} y \\ 0 \end{pmatrix}, \qquad \mathbf{P} := (p_k(x_i))_{i,k=1}^{m,n}. \tag{33}$$

If the points $\{x_i : i = 1, \dots, m\}$, $m \geq \dim\left(\Pi_{\nu-1}(\mathbb{R}^d)\right)$ are $\Pi_{\nu-1}(\mathbb{R}^d)$-*unisolvent*, i.e., the zero polynomial is the only polynomial from $\Pi_{\nu-1}(\mathbb{R}^d)$ that vanishes on all of them, then the linear system of equations (33) has a unique solution. To verify that the coefficient matrix in (33) is indeed invertible, consider the corresponding homogeneous system. Then the second system of equations $\mathbf{P}^\mathrm{T}\alpha = 0$ means that α satisfies (39). Multiplying the first system of equations by α^T gives $0 = \alpha^\mathrm{T}\mathbf{K}\alpha + (\mathbf{P}^\mathrm{T}\alpha)^\mathrm{T}\beta = \alpha^\mathrm{T}\mathbf{K}\alpha$. By the definition of conditionally positive definite functions of order ν, this is only possible if $\alpha = 0$. But then $\mathbf{P}\beta = 0$. Since the points $\{x_i : i = 1, \dots, m\}$ are $\Pi_{\nu-1}(\mathbb{R}^d)$-unisolvent, this implies that $\beta = 0$.

The interpolation by radial basis functions having (conditionally) positive definite kernels was examined, including fast evaluation techniques for the interpolating function f, by many authors; for an overview see, e.g., the books of Buhmann 2003 [18], Wendland [114], and Fasshauer 2007 [39].

Relation to Kriging The interpolation results can be derived in another way by so-called kriging. Kriging is a group of geostatistical techniques to interpolate the unknown value of a random field from observations of its value at nearby locations. Based on the pioneering work of Krige [59] on the plotter of the distance-weighted average gold grades at the Witwatersrand reef in South Africa, the French mathematician Matheron [70] developed its theoretical foundations. Let $S(x)$ denote a random field such that the expectation value fulfills $E(S(X)) = 0$ which is the setting in the so-called simple kriging. Let $K(x_i, x_j) := \mathrm{Cov}(S(x_i), S(x_j))$ and $\mathbf{K} := \left(K(x_i, x_j)\right)_{i,j=1}^m$. The aim is to approximate the value $S(x_0)$ from observations $S(x_i) = y_i$, $i = 1, \dots, m$ by the kriging estimator

$$\hat{S}(x_0) = \sum_{i=1}^m \omega_i(x_0) S(x_i)$$

in such a way that the variance of the error is minimal, i.e.,

$$\begin{aligned} \mathrm{Var}(\hat{S} - S) &= \mathrm{Var}(\hat{S}) + \mathrm{Var}(S) - 2\mathrm{Cov}(S, \hat{S}) \\ &= \sum_{i=1}^m \sum_{j=1}^m \omega_i(x_0)\omega_j(\hat{x}) K(x_i, x_j) \\ &\quad -2\sum_{i=1}^m \omega_i(x_0) K(x_0, x_i) + \mathrm{Var}(S) \;\to\; \min_{\omega(x_0)}. \end{aligned}$$

Setting the gradient to zero, the minimizer $\omega^* = \left(\omega_1^*(x_0), \ldots, \omega_m^*(x_0)\right)^{\mathrm{T}}$ is given by the solution of the following linear system of equations:

$$\mathbf{K}\omega = (K(x_0, x_i))_{i=1}^m .$$

In case $\mathbf{K}$ is invertible, we get

$$S(x_0) = \langle y, \mathbf{K}^{-1} (K(x_0, x_i))_{i=1}^m \rangle = \langle \mathbf{K}^{-1} y, (K(x_0, x_i))_{i=1}^m \rangle .$$

Supposing the same values $K(x_i, x_j)$ as in the interpolation task, this is exactly the same value as $f(x_0)$ from the radial basis interpolation problem (32).

Least Squares Classification and Regression

Also in the feature space, least squares classification and regression can be treated by the same model. So-called Least Squares Support Vector Classifiers were introduced by Suykens and Vandevalle [99] while least squares regression was also considered within *regularization network* approaches, e.g., by Evgeniou et al. [37] and Wahba [112]. The least squares model in the feature space is

LS classification/regression in feature space (Primal problem)

$$\tfrac{1}{2}\|w\|_{\ell_2(I)}^2 + \tfrac{C}{2} \sum_{i=1}^m (\langle w, \Phi(x_i) \rangle_{\ell_2(I)} + b - y_i)^2 \;\to\; \min_{w,b} .$$

and becomes in the case that the feature map is related with a Mercer kernel K a problem in a RKHS $\mathcal{H}_K$:

LS classification/regression *in RKHS* (Primal problem)

$$\tfrac{1}{2}\|f\|_{\mathcal{H}_K} + \tfrac{C}{2} \sum_{i=1}^m (f(x_i) + b - y_i)^2 \;\to\; \min_{w,b} .$$

Setting gradients to zero, one can try to solve this primal problem via a linear system of equations (19) with $\mathbf{X} := (\Phi(x_1) \ldots \Phi(x_m))$. However, one has to compute with $\mathbf{X}\mathbf{X}^{\mathrm{T}}$ here which is only possible if the feature space is finite dimensional. In contrast, the dual approach leads to the linear system of equations (20) which involves only the kernel matrix $\mathbf{K} = \mathbf{X}^{\mathrm{T}}\mathbf{X}$ from (26). Knowing the dual variable α^*, the optimal function f_{w^*} can be computed using (21) with the feature space modification as

$$f_{w^*}(x) = \sum_{i=1}^m \alpha_i^* \langle \Phi(x_i), \Phi(x) \rangle \sum_{i=1}^m \alpha_i^* K(x_i, x).$$

In general there is no sparse representation with respect to the vectors x_i. For more information on least squares kernel methods, the reader may consult the book of Suykens et al. [98]. Figure 7, right, shows a least squares SVM function for the data in Fig. 2.

Other Models

There are numerous models related to the classification and regression models of the previous subsections. A simple classification model which uses only the hinge loss function without penalizing the weight was proposed by Bennet and Mangasarian [7]:

$$\sum_{i=1}^{m} L_h\left(y_i, \langle w, x_i\rangle + b\right) \;\to\; \min_{w,b}.$$

This approach is called *robust linear programming* (RLP) and requires only linear programming methods. Note that the authors weighted the training errors by $1/n_{\pm}$, where $n_{\pm} := |\{i; y_i = \pm 1\}|$. The *linear SV soft margin classifier* adds just the penalizer $\frac{\lambda}{2}\|w\|_2^2$ with $\lambda = 1/C$, $0 < C < \infty$ to the RLP term which leads to quadratic programming. Alternatively, one can add instead the ℓ_2-norm the ℓ_1-norm of the weight as it was done by Bradley and Mangasarian [16]:

$$\sum_{i=1}^{m} L_h\left(y_i, \langle w, x_i\rangle + b\right) + \lambda\|w\|_1 \;\to\; \min_{w,b}.$$

As in (30), the ℓ_1 penalizer enforces the sparsity of the solution vector w^*. Note that the sparsity of w^* itself and not a sparse representation of w^* as linear combination of the support vectors x_i is announced here. The ℓ_1-penalizer was introduced in the statistical context of linear regression in conjunction with the least squares loss function by Tibshirani [101] and is called "LASSO" (*L*east *A*bsolute *S*hrinkage and *S*election *O*perator):

$$\sum_{i=1}^{m} L_{\mathrm{lsq}}\left(y_i, \langle w, x_i\rangle + b\right) + \lambda\|w\|_1 \;\to\; \min_{w,b}$$

As emphasized in section "Support Vector Regression," the ℓ_1-norm is more or less a replacement for the ℓ_0-norm to make problems computable. Other substitutes of the ℓ_0-norm are possible, e.g., $\|w\|_{\ell_0} \approx \sum_{j=1}^{d}(1 - e^{\nu|w_j|})$, $\nu > 0$ which gives

$$\sum_{i=1}^{m} L_h\left(y_i, \langle w, x_i\rangle + b\right) + \lambda \sum_{j=1}^{d}(1 - e^{\nu|w_j|}) \;\to\; \min_{w,b}.$$

This is a nonconvex model and was proposed by Bradley and Mangasarian [16] and by Weston et al. [115] as FSV (*F*eatures *S*election concav*V*e). Numerical solution methods via *successive linearization algorithms* and *difference of convex functions* algorithms by Tao and An [100] were applied.

Further, one can *couple several penalizers* and *generalize the models to the feature space* to obtain nonlinear classifiers as it was done, e.g., by Neumann et

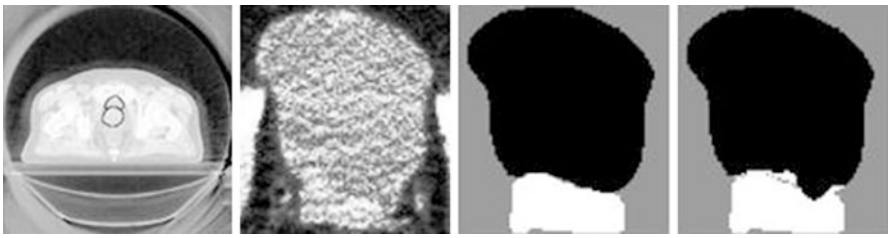


Fig. 8 From left to right: (i) Sample CT slice from a three-dimensional scan of the data set with contours of bladder and prostate. (ii) A zoom within the region of interest shows that the organs are very difficult to distinguish visually. (iii) Manual classification by an expert as "accepted truth." (iv) A classification result: The images are filtered by a three-dimensional steerable pyramid filter bank with 16 angular orientations and four decomposition levels. Then local histograms are built for the filter responses with ten bins per channel. Including the original grey values, this results in 650 features per image voxel which are used for classification by the "ℓ_2-ℓ_0-SV" machine

al. [75]. Figure 8 shows an example for the binary classification of specific organs in CT scans with a so-called "ℓ_2-ℓ_0-SV" machine taken from the above paper, where more information can be found. In particular, a χ^2 kernel was used here.

Multi-class Classification and Multitask Learning

So far only binary classification was considered. Assume now that one wants to learn $K > 2$ classes. Figure 9 shows a typical example of 12 classes for the classification of Mammals; see Amit et al. [2].

Some attempts to extend the binary case to multi-classes were achieved by adding constraints for every class; see Weston and Watkins [116] and Vapnik [105]. In case of many classes, this approach often results in quadratic problems which are hard to solve and difficult to store. In the following, two general approaches to handle multiple classes are presented, namely, with:

(i) Vector-valued binary class labeling,
(ii) K class labeling.

The dominating approach for solving multi-class problems using SVMs is based on reducing a single multi-class problem to multiple binary problems. For instance, a common method is to build a set of binary classifiers, where each classifier distinguishes between one of the labels and the rest. This so-called *one-versus-all* classifier cannot capture correlations between different classes since it breaks the multi-class problem into independent binary problems. More general, one can assign to each class a *vector-valued binary class label* $(y^{(1)}, \ldots, y^{(\kappa)})^{\mathrm{T}} \in \{-1, 1\}^{\kappa}$ and use a classifier based on

$$F(x) := \left(\operatorname{sgn}(\langle w^{(k)}, x\rangle + b^{(k)})\right)_{k=1}^{\kappa}.$$

Fig. 9 Classification of mammal images: 12 from 72 classes of animals which were used for classification in Amit et al. [2]. Typically these classes share common characteristics as in the different rows above (deers, canines, felines, rodents), e.g., the texture or shape

For example, in the one-versus-all method, the classes can be labeled by $\{(-1 + 2\delta_{r,k})_{k=1}^{K} : r = 1, \ldots, K\}$, i.e., $\kappa = K$, and the assignment of x to a class can be made according to the shortest Hamming distance of $F(x)$ from these class labels. In the one-versus-all example, there was $\kappa = K$. More sophisticated methods use values $\kappa > K$ and error-correcting output codes as Dietterich and Bakiri [32]. Note that 2^κ different labels are in general possible with binary vectors of length κ which is an upper bound for the number of classes that could be learned. In the learning process, one can obtain $w^{(k)} \in \mathbb{R}^m$ and $b^{(k)} \in \mathbb{R}$ by solving, e.g.,

$$\frac{1}{2}\sum_{k=1}^{\kappa} \|w^{(k)}\|^2 + \sum_{k=1}^{\kappa} C_k \sum_{i=1}^{m} L\left(y_i, \langle w^{(k)}, x_i\rangle + b^{(k)}\right) \to \min_{w^{(k)}, b^{(k)}}, \tag{34}$$

where L is some loss function. Note that this problem can be decoupled with respect to k. Let $W := (w^{(1)} \ldots w^{(\kappa)}) \in \mathbb{R}^{d,\kappa}$ be the weight matrix. Then the first sum in (34)

coincides with the squared *Frobenius norm* of W defined by

$$\|W\|_F^2 := \sum_{k=1}^{\kappa} \sum_{i=1}^{d} (w_i^{(k)})^2.$$

Let us consider the second labeling approach. Here one assumes that each class label is an integer from $\mathcal{Y} := \{1, \ldots, K\}$. As before, one aims to learn weight vectors $w^{(k)}$, $k = 1, \ldots, K$ (the intercept is neglected for simplicity here). The classifier is given by

$$F_W(x) := \operatorname*{argmax}_{k=1,\ldots,K} \langle w^{(k)}, x \rangle.$$

A training sample (x_i, y_i) is correctly classified by this classifier if

$$\langle w^{(y_i)}, x_i \rangle \geq \langle w^{(k)}, x_i \rangle + 1, \quad \forall k = 1, \ldots, K,\ k \neq y_i.$$

Without adding 1 at the left-hand side of the inequality, correct classification is still attained if there is strong inequality for $k \neq y_i$. This motivates to learn the weight vectors by solving the minimization problem

$$\tfrac{1}{2}\|W\|_F^2 \to \min_W \text{ subject to } \langle w^{(y_i)}, x_i \rangle + \delta_{y_i,k} - \langle w^{(k)}, x_i \rangle \geq 1, \quad \forall k = 1, \ldots, K \text{ and } i = 1, \ldots, m.$$

After introducing slack variables to relax the constraints, one gets

$$\tfrac{1}{2}\|W\|_F^2 + C \sum_{i=1}^{m} \xi_i \to \min_{W,\xi_i} \text{subject to} \langle w^{(y_i)}, x_i \rangle + \delta_{y_i,k} - \langle w^{(k)}, x_i \rangle \geq 1 - \xi_i, \quad \forall k = 1, \ldots, K \text{ and } i = 1, \ldots, m.$$

This can be rewritten as the following unconstrained problem:

$$\frac{1}{2}\|W\|_F^2 + C \sum_{i=1}^{m} l_h \left(\langle w^{(y_i)}, x_i \rangle + \max_k (\langle w^{(k)}, x_i \rangle - \delta_{y_i,k}) \right) \to \min_W.$$

In this functional the learning tasks are coupled in the loss function.

In general the aim of *multitask learning* is to learn data that are common across multiple related supervised learning tasks, i.e., to facilitate "cooperative" learning. Recently, multitask learning has received attention in various applications; see the paper of Caruana [21]. Learning of vector-valued functions in RKHSs was considered, e.g., by Micchelli and Pontil [72]. Inspired by the "sparseness" models in section "Other Models" which focus on the sparsity of the weight vector w, one can ask for similar approaches for a weight matrix W. As a counterpart of the ℓ_0-norm of a weight vector, there can serve a low rank of the weight matrix. But

as in ℓ_0-penalized minimization problems, such problems are computationally not manageable. A remedy is to replace the low-rank condition by demanding a small *trace norm* or *nuclear norm* of W defined by

$$\|W\|_* := \sum_j \sigma_j ,$$

where σ_j are the singular values of W. Then a minimization problem to learn the weight matrix reads, for example, as

$$\frac{1}{2}\|W\|_* + C \sum_{k=1}^{K} \sum_{i=1}^{m} L\left(y_i, \langle w^{(k)}, x_i \rangle\right), \to \min_W$$

where L is mainly the least squares loss function. Such models were considered by Amit et al. [2], Obozinski et al. [76], and Pong et al. [82]. Other approaches use the norm

$$\|W\|_{2,1} := \sum_{j=1}^{d} \left\| \left(w_j^{(k)}\right)_{k=1}^{K} \right\|$$

which favors a small number of nonzero rows in W instead of the trace norm; see Argyriou et al. [4] and Obozinski et al. [76]. Another interesting model was proposed by Argyriou et al. [4] and learns in addition to a weight matrix an orthogonal matrix $U \in \mathcal{O}$ by minimizing

$$\|W\|_{2,1} + C \sum_{k=1}^{K} \sum_{i=1}^{m} L\left(y_i, \langle w^{(k)}, U x_i \rangle\right). \to \min_{W, U \in \mathcal{O}} .$$

The numerical solution of multitask problems which are convex but non-smooth requires sophisticated techniques. The trace norm minimization problem can be, for example, reformulated as a semi-definite program (SDP), and then existing SDP solvers can be used as long as the size of the problem is moderate; see the papers of Fazel et al. [40] and Srebro et al. [91]. A smooth, but nonconvex reformulation of the problem and a subsequent solution by a conjugate gradient or alternating minimization method was proposed, e.g., by Weimer et al. [113]. Accelerated proximal gradient methods (multistep methods) and Bregman iterative methods were applied in the papers of Lu et al. [66], Ma et al. [67], Cai et al. [20], and Toh and Yun [103]. A new primal-dual reformulation of the problem in conjunction with a gradient projection method to solve the reduced dual problem was given by Pong et al. [82].

Applications of SVMs

SVMs have been applied to many real-world problems. Some applications were already sketched in the previous subsections. Very often SVMs are used in connection with other techniques, in particular feature extraction/selection methods to specify the input domain.

A non-exhaustive list of SVM applications includes *text categorization* (see Joachims [53] and Leopold and Kinderman [64]), *hand-written character recognition* (see LeCun et al. [62]), *texture and image classification* (see Chapelle et al. [23]), *protein homology detection* (see Jaakkola and Haussler [51]), *gene expression* (see Brown et al. [17]), *medical diagnostics* (see Strauss et al. [96]), and *pedestrian and face detection* (see Osuna et al. [77], Viola and Jones [110]).

This subsection describes only two applications of SVM classification and shows how the necessary design choices can be made. In particular, one has to choose an appropriate SVM kernel for the given application. Default options are Gaussians or polynomial kernels, and the corresponding SVMs often already outperform other classification methods. Even for such parameterized families of kernels, one has to specify the parameters like the standard deviation of the Gaussian or the degree of the polynomial. In the Gaussian case, a good choice of the standard deviation in the classification problem is the distance between closest points within different classes. In the absence of reliable criteria, one could use a validation set or cross-validation to determine useful parameters. Various applications require more elaborate kernels which implicitly describe the feature space.

Handwritten Digit Recognition The problem of handwritten digit recognition was the first real-world task on which SVMs were successfully tested. The results are reported in detail in the book of Vapnik [105]. This SVM application was so interesting because other algorithms incorporating prior knowledge on the USPS database have been designed. The fact that SVMs perform better than these specific systems without using prior detailed information is remarkable; see [62].

Different SVM models have been tested on two databases:

- United States Postal Service (USPS): 7,291 training and 2,007 test patterns of the numbers $0, \ldots, 9$, represented by 16×16 gray level matrices; see Fig. 10.
- National Institute of Standard and Technology (NIST): 60,000 training and 10,000 test patterns, represented by 20×20 gray-level matrices.

In the following, the results for the USPS database are considered.

For constructing the decision rules, SVMs with polynomial and Gaussian kernels were used:

$$K(x,t) := (\langle x,t\rangle/256)^n\,, \quad K(x,t) := e^{-\|x-t\|^2/(256\sigma^2)}.$$

Fig. 10 Examples of patterns from the USPS database; see [105]

The overall machine consists of 10 classifiers, each one separating one class from the rest (one-versus-all classifier). Then the 10-class classification was done by choosing the class with the largest output number.

All types of SVMs demonstrated approximately the same performance shown in the following tables, cf. [105]. The tables contain the parameters for the hard margin machines, the corresponding performance, and the average (over one classifier) number of support vectors. Moreover, it was observed that the different types of SVMs use approximately the same set of support vectors.

Degree n	1	2	3	4	5	6
Error	8.9	4.7	4.0	4.2	4.5	4.5
Number of SV	282	237	274	321	374	422

Results for SVM classification with polynomial kernels

σ	4.0	1.5	0.3	0.25	0.2	0.1
Error	5.3	4.9	4.2	4.3	4.5	4.6
Number of SV	266	237	274	321	374	422

Results for SVM classification with Gaussian kernels

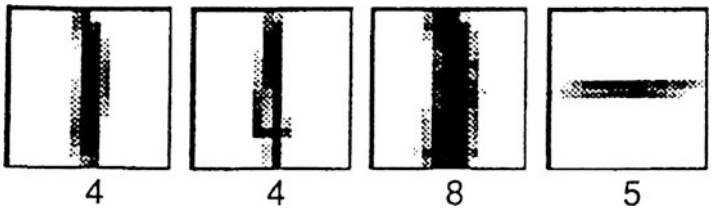


Fig. 11 Labeled USPS examples of training errors for the SVM with second degree polynomial kernel; see [105]

Finally, it is worth to mention that the *training data* are not *linearly* separable; $\approx 5\,\%$ of these data were misclassified by a linear learning machine. For a degree 2 polynomial kernel, only the 4 examples in Fig. 11 were misclassified. For polynomials of degree 3, the training data are separable. The number of support vectors increases only slowly with the degree of the polynomial.

Color Image Recognition Image recognition is another area where SVMs were successfully applied. Chapelle et al. [23] have reported their SVM classification results for color image recognition. The database was a subset (Corel14) of the Corel Stock Photo Collection consisting of 1,400 photos associated with 14 categories. Each category was split into 2/3 for training and 1/3 for testing. Again the one-versus-all classifier was applied.

The images were not used themselves as inputs, but each image was associated to its color histogram. Since each color is a point in a three-dimensional vector space and the number of bins per color was fixed at 16, the dimension of such a histogram (and thus of the feature space) is $d = 16^3$. Note that low-level features like histograms have the advantage that they are invariant with respect to many operations and allow the comparison of images of different sizes. Of course, local high-level image features like edges are not captured by low-level features. Chapelle and co-workers [23] have used both the RGB (red, green, blue) and the HSV/HSB (hue, saturation, value/hue, saturation, brightness) histogram representation. Note that HSV arranges the geometry of RGB in an attempt to be more perceptually relevant. As kernels they have used

$$K(x,t) := e^{-\mathrm{dist}(x,t)/\sigma^2},$$

where dist denotes a measure of similarity in the feature space which has to be determined. For histograms, the χ^2 function

$$\mathrm{dist}(x,t) := \sum_{i=1}^{d} \frac{(x_i - t_i)^2}{x_i + t_i}$$

is accepted as an appropriate distance measure. It is not clear if the corresponding kernel is a Mercer kernel. For the distances $\mathrm{dist}_p(x,t) := \|x-t\|_p^p$, $p = 1, 2$ this is the case.

As can be seen in the following table, the SVM with the χ^2 and the ℓ_1 distance perform similarly and significantly better than the SVM with the squared ℓ_2 distance. Therefore, the Gaussian kernel is not the best choice here. RGB- and HSV-based methods perform similarly.

	Linear	Degree 2 poly	χ^2	ℓ_1	Gaussian
RGB	42.1	33.6	28.8	14.7	14.7
HSV	36.3	35.3	30.5	14.5	14.7

Error rates (percent) using different SVM kernels

For comparision, Chapelle and co-workers conducted some experiments of color image histogram (HSV-based) classifications with the K-nearest neighbor algorithm with χ^2 and ℓ_2. Here $K = 1$ gives the best result presented in the following table:

χ^2	ℓ_2
26.5	47.7

Error rates (percent) with k-nearest neighbor algorithm

The χ^2-based SVM is roughly twice as good as the χ^2-based k-nearest neighbor technique.

4 Survey of Mathematical Analysis of Methods

Reproducing Kernel Hilbert Spaces

General theory For simplicity, let $\mathcal{X} \subset \mathbb{R}^d$ be a compact set throughout this subsection. Moreover, only spaces of real-valued functions are considered. Let $C(\mathcal{X})$ denote the set of continuous functions on $\mathcal{X}$. Together with the norm

$$\|f\|_{C(\mathcal{X})} = \sup\{|f(x)| : x \in \mathcal{X}\}$$

this becomes a Banach space. Further, let $L_2(\mathcal{X})$ be the Hilbert space of (equivalence classes) quadratic integrable, real-valued functions on $\mathcal{X}$ with inner product and norm given by

$$\langle f, g\rangle_{L_2} := \int_{\mathcal{X}} f(x)g(x)\,dx, \quad \|f\|_{L_2} = \left(\int_{\mathcal{X}} f(x)^2\,dx\right)^{1/2}.$$

Since $\mathcal{X}$ is compact, the space $C(\mathcal{X})$ is continuously embedded into $L_2(\mathcal{X})$ which means that $\|f\|_{L_2(\mathcal{X})} \le C\|f\|_{C(\mathcal{X})}$ for all $f \in C(\mathcal{X})$. A function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$

is *symmetric*, if $K(x,t) = K(t,x)$ for all $x,t \in \mathcal{X}$. With a symmetric function $K \in L_2(\mathcal{X} \times \mathcal{X})$, one can associate an integral operator $T_K : L_2(\mathcal{X}) \to L_2(\mathcal{X})$ by

$$T_K f(t) := \int_{\mathcal{X}} K(x,t) f(x)\, dx.$$

This operator is a compact and self-adjoint operator, and K is called its *kernel*. The following spectral theorem holds true for compact, self-adjoint operators, i.e., in particular for T_K.

Theorem 1 (Spectral theorem for compact, self-adjoint operators). *Let T be a compact, self-adjoint operator on the Hilbert space $\mathcal{H}$. Then there exists a countable (possibly finite) orthonormal system $\{\psi_i : i \in I\}$ and a zero sequence $(\lambda_i)_{i\in I}$, $\lambda_i \in \mathbb{R}\backslash\{0\}$ such that*

$$\mathcal{H} = \ker T \oplus \overline{\operatorname{span}\{\psi_i : i \in I\}}$$

and

$$Tf = \sum_{j\in I} \lambda_j \langle f, \psi_j \rangle_{\mathcal{H}}\, \psi_j \quad \forall f \in \mathcal{H}. \tag{35}$$

The numbers λ_j are the nonzero eigenvalues of T and ψ_j are the corresponding eigenfunctions. If T is a positive operator, *i.e.,*

$$\langle Tf, f \rangle_{\mathcal{H}} = \int_{\mathcal{X}} \int_{\mathcal{X}} K(x,t) f(x) f(t)\, dx dt \geq 0 \quad \forall f \in \mathcal{H},$$

then the values λ_j are positive.

Consider the special operator T_K for a symmetric kernel $K \in L_2(\mathcal{X} \times \mathcal{X})$. Using the L_2-orthonormal eigenfunctions $\{\psi_i : i \in I\}$ of T_K, one can also expand the kernel itself as

$$K(x,t) = \sum_{j\in I} \lambda_j \psi_j(x) \psi_j(t),$$

where the sum converges as those in (35) *in general only in* $L_2(\mathcal{X} \times \mathcal{X})$. One can tighten the statement if K is continuous and symmetric. Then $T_K : C(\mathcal{X}) \to C(\mathcal{X})$ is a compact operator on the Pre-Hilbert spaces $C(\mathcal{X})$ equipped with the L_2-norm into itself, and the functions ψ_j are continuous. If $f \in C(\mathcal{X})$, then the right-hand side in (35) converges absolutely and uniformly. To prove such a convergence result also for the kernel expansion, one needs moreover that the operator T_K is positive. Unfortunately, it is not true that a positive kernel K implies a positive operator T_K. There is another criterion which will be introduced in the following. A matrix

$\mathbf{K} \in \mathbb{R}^{m,m}$ is called *positive semi-definite* if

$$\alpha^{\mathrm{T}} \mathbf{K} \alpha \geq 0 \quad \forall \alpha \in \mathbb{R}^m$$

and *positive definite* if strong inequality holds true for all $\alpha \neq 0$. A symmetric kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is *positive (semi)-definite* if the matrix $\mathbf{K} := \big(K(x_i, x_j)\big)_{i,j=1}^m$ is positive (semi)-definite for all finite sets $\{x_1, \ldots, x_m\} \subset \mathcal{X}$. Now a symmetric kernel $K \in C(\mathcal{X} \times \mathcal{X})$ is positive semi-definite if and only if the corresponding integral operator T_K is positive.

Theorem 2 (Mercer's theorem). *Let $K \in C(\mathcal{X} \times \mathcal{X})$ be a continuous, symmetric, and positive semi-definite function with corresponding integral operator T_K. Then K can be expanded into an absolutely and uniformly convergent series in terms of T_K's orthonormal eigenfunctions ψ_j and the associated eigenvalues $\lambda_j > 0$ as follows:*

$$K(x, t) = \sum_{j \in I} \lambda_j \psi_j(x) \psi_j(t). \tag{36}$$

Moreover, if K is positive definite, then $\{\psi_i : i \in I\}$ form an orthonormal basis of $L_2(\mathcal{X})$.

A continuous, symmetric, positive semi-definite kernel is called a *Mercer kernel*. Mercer kernels are closely related to so-called reproducing kernel Hilbert spaces.

Let $\mathcal{H}$ be a real Hilbert space of functions $f : \mathcal{X} \to \mathbb{R}$. A function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a *reproducing kernel* of $\mathcal{H}$ if

(H1) $K_t := K(\cdot, t) \in \mathcal{H} \quad \forall t \in \mathcal{X}$,
(H2) $f(t) = \langle f, K_t \rangle_{\mathcal{H}} \quad \forall f \in \mathcal{H}$ and $\forall t \in \mathcal{X}$ (*Reproducing Property*).

In particular, property (H2) implies for $f := \sum_{i=1}^m \alpha_i K_{x_i}$ and $g := \sum_{j=1}^n \beta_j K_{x_j}$ that

$$\langle f, g \rangle_{\mathcal{H}} := \sum_{i=1}^m \sum_{j=1}^n \alpha_i \beta_j K(x_i, t_j), \quad \|f\|_{\mathcal{H}}^2 = \sum_{i=1}^m \sum_{i,j=1}^m \alpha_i \alpha_j K(x_i, t_j) = \alpha^{\mathrm{T}} \mathbf{K} \alpha, \tag{37}$$

where $\alpha = (\alpha_1, \ldots, \alpha_m)^{\mathrm{T}}$ and $\mathbf{K} := \big(K(x_i, x_j)\big)_{i,j=1}^m$. If such a kernel exists for $\mathcal{H}$, then it is uniquely determined. A Hilbert space which exhibits a reproducing kernel is called *reproducing kernel Hilbert space* (RKHS); write $\mathcal{H} = \mathcal{H}_K$ to emphasize the relation with the kernel. In $\mathcal{H}_K$, the set of all finite linear combinations of K_t, $t \in \mathcal{X}$ is dense, i.e.,

$$\mathcal{H}_K = \overline{\operatorname{span}\{K_t : t \in \mathcal{X}\}}. \tag{38}$$

Moreover, the kernel K of a RKHS must be a symmetric, positive semi-definite function; see Wendland [114]. Finally, based on the Riesz representation theorem, another characterization of RKHSs can be given. It can be shown that a Hilbert space $\mathcal{H}$ is a RKHS if and only if the point evaluation functionals $\delta_x : \mathcal{H} \to \mathbb{R}$ determined by $\delta_x f := f(x)$ are continuous on $\mathcal{H}$, i.e.,

$$|f(x)| \leq C \|f\|_{\mathcal{H}} \quad \forall f \in \mathcal{H}.$$

Conversely, by the following theorem, any Mercer kernel gives rise to a RKHS.

Theorem 3. *Let $K \in C(\mathcal{X} \times \mathcal{X})$ be a continuous, symmetric, and positive semi-definite function. Then there exists a unique Hilbert space $\mathcal{H}_K$ of functions on $\mathcal{X}$ which is a RKHS with kernel K. The space $\mathcal{H}_K$ consists of continuous functions on $\mathcal{X}$, and the embedding operator $\iota_K : \mathcal{H}_K(\mathcal{X}) \to C(\mathcal{X})\big(\to L_2(\mathcal{X})\big)$ is continuous.*

Proof. 1. First, one constructs a Hilbert space which fulfills (H1) and (H2). By (H1), the space $\mathcal{H}_K$ has to contains all functions K_t, $t \in \mathcal{X}$ and since the space is linear also their finite linear combinations. Therefore, define

$$\mathcal{H}_0 := \operatorname{span}\{K_t : t \in \mathcal{X}\}.$$

According to (37), this space can be equipped with the inner product and corresponding norm

$$\langle f, g \rangle_{\mathcal{H}_0} := \sum_{i=1}^{m} \sum_{j=1}^{n} \alpha_i \beta_j K(x_i, t_j), \quad \|f\|_{\mathcal{H}_0}^2 = \alpha^{\mathrm{T}} \mathbf{K} \alpha.$$

It can easily be checked that this is indeed an inner product. In particular $\|f\|_{\mathcal{H}_0} = 0$ for some $f = \sum_{i=1}^{m} \alpha_i K_{x_i}$ implies that $f(t) = \sum_{i=1}^{m} \alpha_i K(t, x_i) = 0$ for all $t \in \mathcal{X}$ by the following argument: set $x_{m+1} := t$. By the positive semi-definiteness of K is follows for any $\epsilon \in \mathbb{R}$ that

$$(\alpha^{\mathrm{T}}, \epsilon) \left(K(x_i, x_j)\right)_{i,j=1}^{m+1} \begin{pmatrix} \alpha \\ \epsilon \end{pmatrix} = \alpha^{\mathrm{T}} \mathbf{K} \alpha + 2\epsilon \sum_{i=1}^{m} \alpha_i K(x_i, t) + \epsilon^2 K(t, t) \geq 0.$$

With $\alpha^{\mathrm{T}} \mathbf{K} \alpha = \|f\|_{\mathcal{H}_0} = 0$, this can be rewritten as

$$\epsilon \left(2 f(t) + \epsilon K(t, t)\right) \geq 0.$$

Since K is positive semi-definite, it holds $K(t, t) \geq 0$. Assume that $f(t) < 0$. Then choosing $0 < \epsilon < -2 f(t)/K(t, t)$ if $K(t, t) > 0$ and $0 < \epsilon$ if $K(t, t) = 0$ leads to a contradiction. Similarly, assuming that $f(t) > 0$ and choosing $-2 f(t)/K(t, t) < \epsilon < 0$ if $K(t, t) > 0$ and $\epsilon < 0$ if $K(t, t) = 0$ gives a contradiction. Thus $f(t) = 0$.

Now one defines $\mathcal{H}_K$ to be the completion of $\mathcal{H}_0$ with the associated norm. This space has the reproducing property (H2) and is therefore a RKHS with kernel K.

2. To prove that $\mathcal{H}_K$ is unique, assume that there exists another Hilbert space $\mathcal{H}$ of functions on $\mathcal{X}$ with kernel K. By (H1) and (38), it is clear that $\mathcal{H}_0$ is a dense subset of $\mathcal{H}$. By (H2) it follows that $\langle f, g\rangle_{\mathcal{H}} = \langle f, g\rangle_{\mathcal{H}_K}$ for all $f, g \in H_0$. Since both $\mathcal{H}$ and $\mathcal{H}_K$ are completions of $\mathcal{H}_0$, the uniqueness follows.
3. Finally, one concludes by the Schwarz inequality that

$$|f(t)| = |\langle f, K_t\rangle_{\mathcal{H}_K}| \le \|f\|_{\mathcal{H}_K}\|K_t\|_{\mathcal{H}_K} = \|f\|_{\mathcal{H}_K}\sqrt{K(t,t)}$$

so that f is continuous since K is continuous. Moreover, $\|f\|_{C(\mathcal{X})} \le C\|f\|_{\mathcal{H}_K}$ with $C := \max_{t\in\mathcal{X}} \sqrt{K(t,t)}$ which means that the embedding ι_K is continuous. □

Since the completion of $\mathcal{H}_0$ is rather abstract, another characterization of $\mathcal{H}_K$ based on Mercer's theorem is useful. Let $\{\psi_i : i \in I\}$ be the L_2-orthonormal eigenfunctions of T_K with corresponding eigenvalues $\lambda_j > 0$ from the Mercer theorem. Then it follows by Schwarz's inequality and Mercer's theorem for $w := (w_i)_{i\in I} \in \ell_2(I)$ that

$$\sum_{i\in I} |w_i\sqrt{\lambda_i}\psi_i(x)| \le \|w\|_{\ell_2}\left(\sum_{i\in I}\lambda_i\psi_i^2(x)\right)^{1/2} = \|w\|_{\ell_2}\sqrt{K(x,x)}$$

so that the series $\sum_{i\in I} w_i\sqrt{\lambda_i}\psi_i(x)$ converges absolutely and uniformly for all $(w_i)_{i\in I} \in \ell_2(I)$. Now another characterization of $\mathcal{H}_K$ can be given.

Corollary 1. *Let $K \in C(\mathcal{X}\times\mathcal{X})$ be a continuous, symmetric, and positive semi-definite kernel with expansion (36). Then the Hilbert space*

$$\mathcal{H} := \left\{\sum_{i\in I} w_i\sqrt{\lambda_i}\psi_i : (w_i)_{i\in I} \in \ell_2(I)\right\}$$

with inner product

$$\langle f, g\rangle_{\mathcal{H}} := \sum_{i\in I} w_i\omega_i = \sum_{j=1}^{\infty}\frac{1}{\lambda_j}\langle f, \psi_j\rangle_{L_2}\langle g, \psi_j\rangle_{L_2}$$

for $f := \sum_{i\in I} w_i\sqrt{\lambda_i}\psi_i$ and $g := \sum_{j\in I}\omega_j\sqrt{\lambda_j}\psi_j$ is the RKHS with kernel K, i.e., $\mathcal{H} = \mathcal{H}_K$. The system $\{\varphi_i := \sqrt{\lambda_i}\psi_i : i \in I\}$ is an orthonormal basis of $\mathcal{H}$.

If K is positive definite, then $\mathcal{H}$ can be also characterized by

$$\mathcal{H} = \{f \in L_2(\mathcal{X}) : \sum_{j=1}^{\infty} \frac{1}{\lambda_j} |\langle f, \psi_j \rangle_{L_2}|^2 < \infty.\}$$

Proof. By the above definition of the inner product, the set $\{\sqrt{\lambda_i}\psi_i : i \in I\}$ is an orthonormal basis of $\mathcal{H}$. The second equality in the definition of the inner product follows by the orthonormality of the ψ_i in L_2.

It remains to show that K fulfills (H1) and (H2). Concerning (H1), it holds $K_t = \sum_{i \in I} \sqrt{\lambda_i}\psi_i(t)\sqrt{\lambda_i}\psi_i$, and since

$$\sum_{i \in I} (\sqrt{\lambda_i}\psi_i(t))^2 = K(t,t) < \infty,$$

it follows that $K_t \in \mathcal{H}$. Using the orthonormal basis property, one can conclude with respect to (H2) that

$$\langle f, K_t \rangle_{\mathcal{H}} = \langle \sum_{j \in I} w_j \sqrt{\lambda_j}\psi_j, \sum_{i \in I} \sqrt{\lambda_i}\psi_i(t)\sqrt{\lambda_i}\psi_i \rangle_{\mathcal{H}} = \sum_{i \in I} w_i \sqrt{\lambda_i}\psi_i(t) = f(t).$$

□

For more information on RKHS, see the book of Berlinet and Thomas-Agnan [8].

Kernels The choice of appropriate kernels for SVMs depend on the application. Default options are Gaussians or polynomial kernels which are described together with some more examples of Mercer kernels below:

1. Let $\mathcal{X} := \{x \in \mathbb{R}^d : \|x\| \leq R\}$ with radius $R > 0$. Then the *dot product kernels*

$$K(x,t) := \sum_{j=0}^{\infty} a_j (x \cdot t)^j, \quad a_j \geq 0, \sum_{j=1}^{\infty} a_j R^{2j} < \infty$$

are Mercer kernels on $\mathcal{X}$. A proof that these kernels are indeed positive semi-definite is given in the book of Cucker and Zhou [29]. A special case appears if $\mathcal{X}$ contains the coordinate vectors e_j, $j = 1, \ldots, d$ and the kernel is $K(x,t) = 1 + x \cdot t$. Note that even in one dimension $d = 1$, this kernel is not positive definite. Here the corresponding RKHS $\mathcal{H}_K$ is the space of linear functions and $\{1, x_1, \ldots, x_d\}$ forms an orthonormal basis of $\mathcal{H}_K$.

The special dot product $K(x,t) := (c + x \cdot t)^n$, $c \geq 0$, $n \in \mathbb{N}$, also known as *polynomial kernel* was introduced in statistical learning theory by Vapnik 1998 [105]. More general dot products were described, e.g., by Smola, Schöllkopf and Müller [89]. See also all-subset kernels and ANOVA kernels in [88].

2. Next, consider *translation invariant kernels*

$$K(x,t) := \kappa(x-t),$$

where $\kappa : \mathbb{R}^d \to \mathbb{R}$ is a continuous function which has to be even, that is, $\kappa(-x) = \kappa(x)$ for all $x \in \mathbb{R}^d$ to ensure that K is symmetric. Let us see if K is a Mercer kernel on $\mathbb{R}^d$ and hence on any subset $\mathcal{X}$ of $\mathbb{R}^d$. First, one knows from Bochner's theorem that K is positive semi-definite if and only if it is the Fourier transform of a finite nonnegative Borel measure on $\mathbb{R}^d$. Let $\kappa \in L_1(\mathbb{R}^d)$. Then, K is positive definite if and only if κ is bounded and its Fourier transform is nonnegative and non-vanishing.

A special example on $\mathbb{R}$ ($d = 1$) is the *spline kernel* K generated by the "hat function" $\kappa(x) := \max\{0, 1 - |x|/2\}$. Its Fourier transform is $\hat{\kappa}(\omega) = 2(\sin\omega/\omega)^2 \geq 0$. Multivariate examples of this form can be constructed by using, e.g., box splines. Spline kernels and corresponding RKHSs were discussed, e.g., by Wahba [112].

3. A widely used class of translation invariant kernels are *kernels associated with radial functions*. A function $\kappa : \mathbb{R}^d \to \mathbb{R}$ is said to be *radial* if there exists a function $k : [0,\infty) \to \mathbb{R}$ such that $\kappa(x) = k(\|x\|^2)$ for all $x \in \mathbb{R}^d$. For radial kernels define

$$K(x,t) := k(\|x-t\|^2).$$

A result of Schoenberg [85] says that K is positive semi-definite on $\mathbb{R}^d$ if and only if the function k is completely monotone on $[0,\infty)$. Recall that k is *completely monotone* on $(0,\infty)$ if $k \in C^\infty(0,\infty)$ and

$$(-1)^l k^{(l)}(r) \geq 0 \quad \forall l \in \mathbb{N}_0 \text{ and } \forall r > 0.$$

The function k is called completely monotone on $[0,\infty)$ if it is in addition in $C[0,\infty)$.

It holds that K is positive definite if and only if *one* of the following conditions is fulfilled

(i) $k(\sqrt{\cdot})$ is completely monotone on $[0,\infty)$ and not constant,
(ii) There exists a finite nonnegative Borel measure ν on $[0,\infty)$ that is not concentrated at zero such that

$$k(r) = \int_0^\infty e^{-r^2 t}\, d\nu(t).$$

The proofs of these results on radial kernels are contained, e.g., in the book of Wendland [114].

For $c > 0$, the kernels K arising from the following radial functions κ are positive definite:

$$\kappa(x) := e^{-\|x\|^2/c^2} \quad \text{Gaussian,}$$
$$\kappa(x) := (c^2 + \|x\|^2)^{-s}, \quad s > 0 \quad \text{inverse multiquadric,}$$

where the positive definiteness of the Gaussian follows from (i) and those of the inverse multiquadric from (ii) with

$$k(r) = \frac{1}{\Gamma(s)} \int_0^\infty t^{s-1} e^{-c^2 t} e^{-r^2 t}\, dt.$$

Positive definite kernels arising from Wendland's radial basis functions with compact support (see Wendland [114]) were applied in SVM classification by Strauss and Steidl [95].

Finally, the following techniques for creating Mercer kernels are remarkable.

Theorem 4. *Let $K_j \in C(\mathcal{X} \times \mathcal{X})$, $j = 1, 2$ be Mercer kernels and p a polynomial with positive coefficients. Then the following functions are also Mercer kernels:*

(i) $K(x,t) := K_1(x,t) + K_2(x,t)$,
(ii) $K(x,t) := K_1(x,t) K_2(x,t)$,
(iii) $K(x,t) := p(K_1(x,t))$,
(iv) $K(x,t) := e^{K_1(x,t)}$.

Beyond the above Mercer kernels other kernels like kernels for text and structured data (strings, trees), diffusion kernels on graphs or kernel incorporating generative information were used in practice; see the book of Shawe-Taylor and Cristianini [88].

Conditionally positive semi-definite radial functions. In connection with radial basis functions, so-called conditionally positive semi-definite functions $\kappa(x) := k(\|x\|^2)$ were applied for approximation tasks. Let $\Pi_{\nu-1}(\mathbb{R}^d)$ denote the space of polynomials on $\mathbb{R}^d$ of degree $< \nu$. This space has dimension $\dim(\Pi_{\nu-1}(\mathbb{R}^d)) = \binom{d+\nu-1}{\nu}$. A continuous radial function $\kappa : \mathbb{R}^d \to \mathbb{R}$ is *conditionally positive semi-definite of order* ν if for all $m \in \mathbb{N}$, all pairwise distinct points $x_1, \ldots, x_m \in \mathbb{R}^d$, and all $\alpha \in \mathbb{R}^m \backslash \{0\}$ satisfying

$$\sum_{i=1}^m \alpha_i\, p(x_i) = 0 \quad \forall\, p \in \Pi_{\nu-1}(\mathbb{R}^d) \tag{39}$$

the relation

$$\alpha^{\mathrm{T}} \mathbf{K} \alpha \geq 0, \qquad \mathbf{K} := \big(\kappa(x_i - x_j)\big)_{i,j=1}^m$$

holds true for all $\alpha \in \mathbb{R}^m$. If equality is attained only for $\alpha = 0$, the function κ is said to be *conditionally positive definite of order* ν.

The following result is due to Micchelli [71]: For $k \in C[0,\infty) \cap C^\infty(0,\infty)$, the function $\kappa(x) := k(\|x\|^2)$ is conditionally positive semi-definite of order ν if and only if $(-1)^\nu k^{(\nu)}$ is completely monotone on $(0,\infty)$. If k is not a polynomial of degree at most ν, then κ is conditionally positive definite of order ν.

Using this result, one can show that the following functions are conditionally positive definite of order ν:

$$\kappa(x) := (-1)^{\lceil s \rceil}(c^2 + \|x\|^2)^s, \quad s > 0, s \notin \mathbb{N}, \ \nu = \lceil s \rceil \quad \text{multiquadric},$$
$$\kappa(x) := (-1)^{\lceil s/2 \rceil}\|x\|^s, \quad s > 0, s \notin 2\mathbb{N}, \ \nu = \lceil s/2 \rceil,$$
$$\kappa(x) := (-1)^{k+1}\|x\|^{2k} \log \|x\|, \quad k \in \mathbb{N}, \ \nu = k+1 \quad \text{thin plate spline}.$$

A relation of a combination of thin plate splines and polynomials to the reproducing kernels of certain RKHSs can be found in Wahba [112].

Quadratic Optimization

This subsection collects the basic material from optimization theory to understand the related parts in the previous Sect. 3, in particular the relation between primal and dual problems in quadratic programming. More on this topic can be found in any book on optimization theory, e.g., in the books of Mangasarian [68] or Spellucci [90].

A (nonlinear) optimization problem in $\mathbb{R}^d$ has the general form

Primal problem (P)
$\theta(x) \ \to \ \min_x \quad \text{subject to} \quad g(x) \le 0, \ h(x) = 0$

where $\theta : \mathbb{R}^d \to \mathbb{R}$ is a real-valued function and $g : \mathbb{R}^d \to \mathbb{R}^m$, $h : \mathbb{R}^d \to \mathbb{R}^p$ are vector-valued functions. In general, only the case $p < d$ is of interest since otherwise one is confronted with the solution of a (nonlinear) system of equations. The region

$$\mathcal{G} := \{x \in \mathbb{R}^d : g(x) \le 0, h(x) = 0\},$$

where the *objective function* θ is defined and where all constraints are satisfied, is called *feasible region*. There are classes of problems (P) which are well examined as convex optimization problems and in particular special classes of convex problems, namely, linear and quadratic problems. Problem (P) is called *convex*, if θ is a convex function and $\mathcal{G}$ is a convex region. Recall, that $x^* \in \mathcal{G}$ is a *local minimizer* of θ in $\mathcal{G}$ if there exists a neighborhood $\mathcal{U}(x^*)$ of x^* such that $\theta(x^*) \le \theta(x)$ for all $x \in \mathcal{U}(x^*) \cap \mathcal{G}$. For convex problems, any local minimizer x^* of θ in $\mathcal{G}$ is also a *global minimizer* of θ in $\mathcal{G}$ and therefore a solution of the minimization problem.

This subsection deals mainly with the following setting which gives rise to a convex optimization problem:

(C1) θ convex and differentiable,
(C2) $g_i, i = 1, \ldots, m$ convex and differentiable,
(C3) $h_j, j = 1, \ldots, p$ affine linear.

Important classes of problems fulfilling (C1)–(C3) are *quadratic programs*, where the objective function is quadratic and the constraints are (affine) linear and *linear programs*, where the objective function is also linear. The constrained optimization problems considered in Sect. 3 are of this kind.

The function $L : \mathbb{R}^d \times \mathbb{R}^m_+ \times \mathbb{R}^p \to \mathbb{R}$ defined by

$$L(x, \alpha, \beta) := \theta(x) + \sum_{i=1}^{m} \alpha_i g_i(x) + \sum_{j=1}^{p} \beta_j h_j(x)$$

is called the *Lagrangian* associated with (P), and the coefficients α_i and β_j are called *Lagrange multipliers*. Recall that $(x^*, \lambda^*) \in \Omega \times \Xi$, $\Omega \subset \mathbb{R}^d$, $\Xi \subset \mathbb{R}^n$ is called a *saddle point* of a function $\Phi : \Omega \times \Xi \to \mathbb{R}$ if

$$\Phi(x^*, \lambda) \leq \Phi(x^*, \lambda^*) \leq \Phi(x, \lambda^*) \quad \forall (x, \lambda) \in \Omega \times \Xi.$$

There is the following close relation between saddle point problems and min-max problems:

Lemma 1. *Let $\Phi : \Omega \times \Xi \to \mathbb{R}$. Then the inequality*

$$\max_{\lambda \in \Xi} \min_{x \in \Omega} \Phi(x, \lambda) \leq \min_{x \in \Omega} \max_{\lambda \in \Xi} \Phi(x, \lambda)$$

holds true supposed that all extreme points exist. Moreover, in this case, the equality

$$\max_{\lambda \in \Xi} \min_{x \in \Omega} \Phi(x, \lambda) = \Phi(x^*, \lambda^*) = \min_{x \in \Omega} \max_{\lambda \in \Xi} \Phi(x, \lambda)$$

is fulfilled if and only if (x^*, λ^*) is a saddle point of Φ.

The solution of (P) is related to the saddle points of its associated Lagrangian as detailed in the following theorem.

Theorem 5. *If $(x^*, (\alpha^*, \beta^*)) \in \mathbb{R}^d \times (\mathbb{R}^m_+ \times \mathbb{R}^p)$ is a saddle point of the Lagrangian associated with the minimization problem* (P)*, i.e.,*

$$L(x^*, \alpha, \beta) \leq L(x^*, \alpha^*, \beta^*) \leq L(x, \alpha^*, \beta^*) \quad \forall x \in \mathbb{R}^d, \forall (\alpha, \beta) \in \mathbb{R}^m_+ \times \mathbb{R}^p,$$

then x^* is a solution of (P). *Assume that the functions θ, g, h satisfy the conditions* (C1)–(C3) *and that g fulfills in addition the following* Slater condition*:*

$$\text{there exists } x_0 \in \Omega \text{ such that } g(x_0) > 0 \text{ and } h(x_0) = 0.$$

Then, if x^* is a solution of (P)*, there exist $\alpha^* \in \mathbb{R}^m_+$ and $\beta^* \in \mathbb{R}^p$ such that $(x^*, (\alpha^*, \beta^*))$ is a saddle point of the associated Lagrangian.*

By the next theorem, the minimizers of (P) can be also described via the following conditions on the Lagrangian: there exist $x^* \in \mathcal{G}$, $\alpha^* \in \mathbb{R}^m_+$ and $\beta^* \in \mathbb{R}^p$ such that

$$\begin{aligned} &\text{(KKT1)} \quad \nabla_x L(x^*, \alpha^*, \beta^*) = 0, \\ &\text{(KKT2)} \quad (\alpha^*)^{\mathrm{T}} g(x^*) = 0, \quad \alpha^* \geq 0. \end{aligned}$$

These conditions were independently established by Karush and Kuhn and Tucker [60] and are mainly called *Kuhn-Tucker conditions*.

Theorem 6. *Let θ, g and h fulfill* (C1)–(C3). *If x^* satisfies* (KKT1)–(KKT2)*, then x^* is a solution of* (P). *Assume that g fulfills in addition the Slater condition. Then, if x^* is a solution of* (P)*, it also fulfills* (KKT1)–(KKT2).

If there are *only equality constraints in* (P), then a solution is determined by

$$\nabla_x L(x^*, \beta^*) = 0, \quad h(x^*) = 0.$$

For the rest of this subsection, assume that (C1)–(C3) and the Slater condition hold true. Let a solution x^* of (P) exist. Then, by Lemma 1 and Theorem 5, there exist α^* and β^* such that

$$L(x^*, \alpha^*, \beta^*) = \max_{\alpha \in \mathbb{R}^m_+, \beta} \min_x L(x, \alpha, \beta).$$

Therefore, one can try to find x^* as follows: for any fixed $(\alpha, \beta) \in \mathbb{R}^m_+ \times \mathbb{R}^p$, compute

$$\hat{x}(\alpha, \beta) := \operatorname{argmin}_x L(x, \alpha, \beta). \tag{40}$$

If θ is uniformly convex, i.e., there exists $\gamma > 0$ such that

$$\mu\theta(x) + (1-\mu)\theta(y) \geq \theta(\mu x + (1-\mu)y) + \mu(1-\mu)\gamma \|x-y\|^2 \ \forall x, y \in \mathbb{R}^d, \mu \in [0,1],$$

then $\hat{x}(\alpha, \beta)$ can be obtained as the unique solution of

$$\nabla_x L(x, \alpha, \beta) = 0.$$

This can be substituted into L which results in $\psi(\alpha,\beta) := L(\hat{x}(\alpha,\beta),\alpha,\beta)$ and α^* and β^* are the solution of

$$\psi(\alpha,\beta) \;\to\; \max_{\alpha,\beta} \quad \text{subject to} \quad \alpha \geq 0.$$

This problem, which is called the *dual problem* of (P) can often be more easily solved than the original problem since one has only simple inequality constraints. However, this approach is only possible if (40) can easily be solved. Then, finally $x^* = \hat{x}(\alpha^*,\beta^*)$.

The objective functions in the primal problems in Sect. 3 are not strictly convex (and consequently also not uniformly convex) since there does not appear the intercept b in these functions. So let us formulate the dual problem with $\psi(x,\alpha,\beta) := L(x,\alpha,\beta)$ as follows:

Dual problem (D)

$$\psi(x,\alpha,\beta) \;\to\; \max_{x,\alpha,\beta} \quad \text{subject to} \quad \nabla_x L(x,\alpha,\beta) = 0,\ \alpha \geq 0.$$

The solutions of the primal and dual problem, i.e., their minimum and maximum, respectively, coincide according to the following theorem of Wolfe [117].

Theorem 7. *Let θ, g and h fulfill* (C1)–(C3) *and the Slater condition. Let x^* be a minimizer of* (P). *Then there exist α^*,β^* such that x^*,α^*,β^* solves the dual problem and*

$$\theta(x^*) = \psi(x^*,\alpha^*,\beta^*).$$

Duality theory can be handled in a more sophisticated way using tools from Perturbation Theory in Convex Analysis; see, e.g., the book of Bonnans and Shapiro [11]. Let us briefly mention the general idea. Let $v : \mathbb{R}^m \to (-\infty,\infty]$ be an extended function, where only extended functions $\not\equiv \infty$ are considered in the following. The *Fenchel conjugate* of v is defined by

$$v^*(\alpha) := \sup_{p\in\mathbb{R}^m} \{\langle \alpha, x\rangle - v(x)\}$$

and the *biconjugate* of v by $v^{**} := (v^*)^*$. In general, the inequality $v^{**}(x) \leq v(x)$ holds true and becomes an equality if and only if v is convex and lower semicontinuous. (Later the inequality is indicated by the fact that one minimizes the primal and maximizes the dual problem.) For convex, lower semicontinuous functions $\theta : \mathbb{R}^d \to (-\infty,\infty]$, $\gamma : \mathbb{R}^m \to (-\infty,\infty]$ and $g : \mathbb{R}^d \to \mathbb{R}^m$ one considers the primal problems

$$(\mathrm{P}_u) \quad v(u) = \inf_{x\in\mathbb{R}^d} \{\theta(x) + \gamma(g(x)+u)\},$$

$$\text{(P)} \quad v(0) = \inf_{x \in \mathbb{R}^d} \{\theta(x) + \gamma(g(x))\}$$

where $u \in \mathbb{R}^m$ is the "perturbation." With $L(x, \alpha) := \theta(x) + \langle g(x), \alpha \rangle$, the dual problem reads

$$\text{(D}_u\text{)} \quad v^{**}(u) = \sup_{\alpha \in \mathbb{R}^m} \{\langle \alpha, u \rangle - \gamma^*(\alpha) + \inf_{x \in \mathbb{R}^d} L(x, \alpha)\},$$

$$\text{(D)} \quad v^{**}(0) = \sup_{\alpha \in \mathbb{R}^m} \{-\gamma^*(\alpha) + \inf_{x \in \mathbb{R}^d} L(x, \alpha)\}.$$

For the special setting with the indicator function

$$\gamma(y) = \iota_{\mathbb{R}^m_-}(y) := \begin{cases} 0 & \text{if } y \le 0, \\ \infty & \text{otherwise} \end{cases}$$

the primal problem (P) is equivalent to

$$\theta(x) \to \min_x \quad \text{subject to} \quad g(x) \le 0$$

and since $\gamma^* = \iota_{\mathbb{R}^m_+}$ the dual problem (D) becomes

$$\sup_{\alpha \in \mathbb{R}^m} \inf_{x \in \mathbb{R}^d} L(x, \alpha) \quad \text{subject to} \quad \alpha \ge 0.$$

Again, if θ and g are convex and differentiable and θ is uniformly convex, then the unique solution $\hat{x}(\alpha)$ of $\nabla_x L(x, \alpha) = 0$ is the solution of the infimum problem, and the dual problem becomes $\sup_{\alpha \in \mathbb{R}^m} L(\hat{x}(\alpha), \alpha)$ subject to $\alpha \ge 0$.

Results from Generalization Theory

There exists a huge amount of results on the generalization abilities of statistical learning methods and in particular of support vector machines. The following subsection can only give a rough impression on the general tasks considered in this field from a simplified mathematical point of view that ignores technicalities, e.g., the definition of the correct measure and function spaces and what measurable in the related context means. Most of the material is borrowed from the book of Steinwart and Christmann [93], where the reader can find a sound mathematical treatment of the topic.

To start with, remember that the aim in Sect. 3 was to find a function $f : \mathcal{X} \to \mathbb{R}$ from samples $Z := \{(x_i, y_i) : i = 1, \ldots, m\}$ such that $f(x)$ is a good prediction of y at x for $(x, y) \in \mathcal{X} \times \mathcal{Y}$. Let P denote an unknown probability distribution on $\mathcal{X} \times \mathcal{Y}$. Then a general assumption is that the data used in training and testing are identically independent distributed (iid) according to P. The *loss function* or *cost function* $L : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ describes the cost of the discrepancy between the

prediction $f(x)$ and the observation y at x. The choice of the loss function depends on the specific learning goal. In the models of this paper, the loss functions depend on x only via $f(x)$ such that one can simply write $L(y, f(x))$. In Sect. 3, the hinge loss function and the least squares loss function were used for classification tasks. Originally, one was interested in the $0/1$ classification loss $L_{0/1} : \mathcal{Y} \times \mathbb{R} \to \{0, 1\}$ defined by

$$L_{0/1}(y, t) := \begin{cases} 0 \text{ if } y = \operatorname{sgn}(t), \\ 1 \text{ otherwise.} \end{cases}$$

To the loss function, there is associated a *risk* which is the expected loss of f:

$$R_{L,P}(f) := \int_{\mathcal{X}\times\mathcal{Y}} L(x, y, f(x))dP(x, y) = \int_{\mathcal{X}} \int_{\mathcal{Y}} L(x, y, f(x))dP(y|x)dP_{\mathcal{X}}.$$

For example, the $0/1$ loss function has the risk

$$R_{L_{0/1},P}(f) = P\left((x, y) \in \mathcal{X} \times \mathcal{Y} : \operatorname{sgn} f(x) \neq y\right).$$

A function f is considered to be "better" the smaller the risk is. Therefore, one is interested in the *minimal risk* or *Bayes risk* defined by

$$R^*_{L,P} := \inf_{f:\mathcal{X}\to\mathbb{R}} R_{L,P}(f), \tag{41}$$

where the infimum is taken over all possible (measurable) functions. However, since the distribution P is unknown, it is impossible to find a minimizer of $R_{L,P}$. In learning tasks one can exploit finite training sets Z of iid data. A learning method on $\mathcal{X} \times \mathcal{Y}$ maps every data set $Z \in (\mathcal{X} \times \mathcal{Y})^m$ to a function $f_Z : \mathcal{X} \to \mathbb{R}$. A learning methods should produce for sufficiently large training sets Z nearly optimal decision functions f_Z with high probability. A measurable learning method is called *L-risk consistent* for P if

$$\lim_{m\to\infty} P^m\left(Z \in (\mathcal{X} \times \mathcal{Y})^m : R_{L,P}(f_Z) \leq R^*_{L,P} + \varepsilon\right) = 1 \quad \forall \varepsilon > 0$$

and *universally L-risk consistent*, if it is L-risk consistent *for all* distributions P on $\mathcal{X} \times \mathcal{Y}$. The first learning method that was proved to be universally consistent was the so-called nearest neighbor method; see Stone [94]. Many uniformly consistent classification and regression methods are presented in the books of Devroye et al. [30] and Györfi [46]. Consistency does not address the *speed of convergence*, i.e., convergence rates. Unfortunately, the *no-free-lunch theorem* of Devroye [31], says that for every learning method there exists a distribution P for which the learning methods cannot produce a "good" decision function in the above sense with an a priori fixed speed of convergence. To obtain uniform convergence rates, one has to pose additional requirements on P.

Instead of the risk one can deal with the *empirical risk* defined by

$$R_{L,Z}(f) := \frac{1}{m} \sum_{i=1}^{m} L\left(x_i, y_i, f(x_i)\right).$$

Then the law of large numbers shows that $R_{L,Z}(f)$ becomes a "good" approximation of $R_{L,P}(f)$ *for a fixed* f if m is "large enough." However finding the minimizer of

$$\inf_{f:\mathcal{X}\to\mathbb{R}} R_{L,Z}(f) \tag{42}$$

does in general not lead to a good approximation of $R^*_{L,P}$. For example, the function which classifies all $x_i \in X$ correctly and is zero elsewhere is a minimizer of the above functional (42) but gives in general a poor approximation of the optimal decision function according to (41). This is an example of so-called overfitting, where the learning method approximates the training data too closely and has poor generalization/prediction properties. One common way to cope with this phenomenon is to choose a smaller set $\mathcal{F}$ of functions, e.g., subsets of continuous functions, which should have good approximation properties. In the SVMs treated in Sect. 3, this set $\mathcal{F}$ was a RKHS $\mathcal{H}_K$. Then one considers the *empirical risk minimization* (ERM)

$$\inf_{f\in\mathcal{F}} R_{L,Z}(f). \tag{43}$$

Let a minimizer f_Z of (43) be somehow "computed." (In this subsection, the question of the existence and uniqueness of a minimizer of the various functionals is not addressed.) Then one is of course interested in the error $R_{L,P}(f_Z) - R^*_{L,P}$. Using the infinite-sample counterpart of the ERM

$$R^*_{L,P,\mathcal{F}} := \inf_{f\in\mathcal{F}} R_{L,P}(f)$$

this error can be splitted as

$$R_{L,P}(f_Z) - R^*_{L,P} = \underbrace{R_{L,P}(f_Z) - R^*_{L,P,\mathcal{F}}}_{\text{sample error}} + \underbrace{R^*_{L,P,\mathcal{F}} - R^*_{L,P}}_{\text{approximation error}}.$$

The first error, called *sample error*, is a probabilistic one since it depends on random samples, while the second error, called *approximation error*, is a deterministic one. Finding a good balance between both errors is sometimes called *bias-variance problem*, where the bias is related to the approximation error and the variance to the sampling error.

Concerning the approximation error, it turns out that for RKHS $\mathcal{F} = \mathcal{H}_K$ on compact metric spaces $\mathcal{X}$ which are dense in $C(\mathcal{X})$ and continuous, P-integrable, so-called Nemitski losses, this error becomes zero; see Corollary 5.29 in Steinwart and Christmann [93]. In particular, this is true for RKHS with the Gaussian kernel and the loss functions considered in Sect. 3. For relations between the approximation error, interpolation spaces, and K-functionals, see the book of Cucker and Zhou [29] and the references therein.

Concerning the sample error, there is a huge amount of results, and this paper can only cover some basic directions. For a survey on recent developments in the statistical analysis of classification methods, see Boucheron et al. [14]. Based on Hoeffding's inequality, the first of such relations goes back to Vapnik and Chervonenkis [107]. See also the books of Tsypkin [104], Vapnik [105], Anthony and Bartlett [3], and Vidyasagar [109]. To get an impression how such estimates look like, two of them from Proposition 6.18 and 6.22 of the book of Steinwart and Christmann [93] are presented in the following. If $\mathcal{F}$ is *finite* and $L(x, y, f(x)) \leq B$, then it holds for all $m \geq 1$ that

$$P^m\left(Z \in (\mathcal{X} \times \mathcal{Y})^m : R_{L,P}(f_Z) - R^*_{L,P,\mathcal{F}} \geq B\sqrt{\frac{2\tau + 2\ln(2|\mathcal{F}|)}{m}}\right) \leq e^{-\tau} \forall \tau > 0.$$

If the function class $\mathcal{F}$ is *infinite*, in particular not countable, one needs some bounds on the "complexity" of $\mathcal{F}$. The most classical of such a "complexity" measure is the *VC dimension* (see Vapnik and Chervonenkis [107]) applied in connection with the $0/1$ loss function. Another possibility is the use of *covering numbers* or its counterpart *entropy numbers* going back to Kolmogorov and Tikhomirov [57]. The *ε-covering number* of a metric set T with metric d is the size of the smallest ε-net of T, i.e.,

$$N(T, d, \varepsilon) := \inf\left\{n \geq 1 : \exists s_1, \ldots, s_n \in T \text{ such that } T \subset \bigcup_{i=1}^{n} B_d(s_i, \varepsilon)\right\},$$

where $B_d(s, \varepsilon)$ is the closed ball with center s and radius ε. For the estimation of covering numbers see Edmunds and Triebel [35] and Pinkus [79]. Then, for compact $\mathcal{F} \subset L_\infty(X)$, one has basically to replace $|\mathcal{F}|$ in the above relation by its covering number:

$$P^m\left(Z \in (\mathcal{X} \times \mathcal{Y})^m : R_{L,P}(f_Z) - R^*_{L,P,\mathcal{F}} \right.$$
$$\left. \geq B\sqrt{\frac{2\tau + 2\ln(2N(\mathcal{F}, \|\cdot\|_\infty, \varepsilon))}{m}} + 4\varepsilon|L|_{M,1}\right) \leq e^{-\tau}$$

for all $\tau > 0$ and for all $\varepsilon > 0$, where ones assumes in addition that $\|f\|_\infty \le M$, $f \in \mathcal{F}$ and that L is locally Lipschitz continuous with constant $|L|_{M,1}$ here.

Next let us turn to the SVM setting, where an additional term comes along with the loss function, namely, one is interested in minimizers of

$$\inf_{f \in \mathcal{H}_K} \left\{ R_{L,Z}(f) + \lambda \|f\|_{\mathcal{H}_K}^2 \right\}, \quad \lambda > 0$$

with a regularization term $\lambda \|f\|_{\mathcal{H}_K}^2$ that penalizes functions f with large RKHS norms. The techniques developed for ERM analysis can be extended to the SVM setting.

First let us mention that under some assumptions on the loss function, which are fulfilled for the setting in Sect. 3, a unique minimizer $f_{Z,\lambda}$ exists and has the form

$$f_{Z,\lambda} = \sum_{i=1}^{m} \alpha_i K(x_i, \cdot).$$

This was established in the *representer theorem* by Kimeldorf and Wahba [56] for special continuous loss functions and generalized, e.g., in Schölkopf et al. [86]. There also exist a representer-like theorems for the minimizer $f_{P,\lambda}$ of the infinite-sample setting

$$\inf_{f \in \mathcal{H}_K} \left\{ R_{L,P}(f) + \lambda \|f\|_{\mathcal{H}_K}^2 \right\}$$

(see Steinwart [92], de Vito [111], and Dinuzzo et al. [33]). One can show for the infinite-sample setting that the error

$$A(\lambda) := \inf_{f \in \mathcal{H}_K} \left\{ R_{L,P}(f) + \lambda \|f\|_{\mathcal{H}_K}^2 \right\} - R_{L,P,\mathcal{H}_K}^*$$

tends to zero as λ goes to zero and that $\lim_{\lambda \to 0} R_{L,P}(f_{P,\lambda}) = R_{L,P,\mathcal{H}_K}^*$. Let us come to the essential question how close $R_{P,\lambda}(f_{Z,\lambda})$ is to $R_{L,P}^*$. Recall that $R_{L,P}^* = R_{L,P,\mathcal{H}_K}^*$ for the above mentioned RKHS. An ERM analysis like estimation has, for example, the form

$$P^m \Bigg(Z \in (\mathcal{X} \times \mathcal{Y})^m : R_{L,P}(f_{Z,\lambda}) + \lambda \|f_{Z,\lambda}\|_{\mathcal{H}_K}^2 - R_{L,P,\mathcal{H}_K}^* \\ \ge A(\lambda) + (\lambda^{-1/2} |L|_{\lambda^{-1/2},1} + 1) \\ \sqrt{\frac{2\tau + 2\ln(2N(B_{\mathcal{H}_K}, \|\cdot\|_\infty, \lambda^{1/2}\varepsilon))}{m}} + 4\varepsilon |L|_{\lambda^{-1/2},1} \Bigg) \le e^{-\tau},$$

for $\tau > 0$, where one assumes that the continuous kernel fulfills $\|K\|_\infty \le 1$, $L(x, y, 0) \le 1$ and $B_{\mathcal{H}}$ is the closed unit ball in $\mathcal{H}$ (see Theorem 6.25 in Steinwart

and Christmann [93] and also Cucker and Smale [28] and Bousquet and Elisseeff [15]). For a certain decay of the covering number $\ln(2N(B_{\mathcal{H}_K}, \|\cdot\|_\infty, \varepsilon))$ in ε and a RKHS for which the approximation error becomes zero, one can then conclude that for zero sequences $(\lambda_m)_{m\geq 1}$ with an additional suitable decay property related to the decay of the covering number, the relation $R_{L,P}(f_{Z,\lambda_m}) \to R^*_{L,P}$ holds true in probability.

The above relations can be further specified for classification and regression tasks with special loss functions. With respect to classification, one can find, for example, upper bounds for the risk in terms of the margin or the number of support vectors. For the $0/1$ loss function, the reader may consult, e.g., the book of Cristianini and Shawe-Taylor [27]. For the hinge loss function and the soft margin SVM with $C = 1/(2\lambda m)$, it holds, for example, that

$$\frac{|I_S|}{m} \geq 2\lambda \|f_{Z,\lambda}\|^2_{\mathcal{H}_K} + R_{L,Z}(f_{Z,\lambda})$$

(see Proposition 8.27 in Steinwart and Christmann [93]). For a suitable zero sequence $(\lambda_m)_{m\geq 1}$ and a RKHS with zero approximation error, the following relation is satisfied:

$$\lim_{m\to\infty} P^m \left(Z \in (\mathcal{X}\times\mathcal{Y})^m : \frac{|\{i : \alpha_i^*(Z) > 0\}|}{m} \geq R^*_{L,P} - \varepsilon \right) = 1, \quad \varepsilon > 0.$$

Finally, let us address the setting, where the risk function defining the learning task is hard to handle numerically. One example is the risk function associated with the $0/1$ loss function. This function is neither continuous nor convex. One remedy is to replace such unpleasant loss functions L by a convex *surrogate* L_{sur} where one has to ensure that the minimizer f_Z in (43) for the surrogate loss fulfills $R_{L,P}(f_Z) \approx R^*_{L,P}$. For the hinge function as surrogate of the $0/1$ loss function, Zhang [119], has proved that

$$R_{L_{0/1},P}(f) - R^*_{L_{0/1},P} \leq R_{L_h,P}(f) - R^*_{L_h,P}$$

for all measurable functions f. Thus, if $R_{L_h,P}(f_Z) - R^*_{L_h,P}$ is small, this follows for the original risk function, too. For a systematical treatment of surrogate loss functions, the reader may consult Chapter 3 in the book of Steinwart and Christmann [93] which was partially inspired by the work of Barlett et al. [6].

5 Numerical Methods

This section concentrates on the support vector machines in Sect. 3. Numerical methods for the other models were always sketched when they were introduced. Support vector machines require finally the minimization of a quadratic functional subject to linear constraints (QP). These minimization problems involve a symmetric, fully populated kernel matrix having the size m of the training set.

Hence, this matrix has in general $m(m+1)/2$ distinct nonzero coefficients one has to work with. Therefore, one has to distinguish between small- to moderate-sized problems, where such a matrix can be stored into the RAM of the computer, and large-sized problems, say, with more than a million training data.

For quadratic programming with *small to moderate data sizes*, there exist various meanwhile standard algorithms. They are implemented in commercial software packages like CPLEX or MOSEK (see also the MATLAB optimization toolbox or in freeware packages like MINOS and LOQO). Among them, the *primal-dual interior point algorithms* belong to the most reliable and accurate techniques. The main idea of interior point methods is to solve the primal and dual problems simultaneously by enforcing the Kuhn-Tucker conditions to iteratively find a feasible solution. The duality gap, i.e., the difference between the minimum of the primal problem and the maximum of the dual problem, is used to determine the quality of the current set of variables and to check whether the stopping criteria are fulfilled. For QP algorithms including recent algorithms for solving large QPs, the reader may consult the new book of Zdenek [118].

The problem of learning *large data sets* was mainly addressed based on so-called "working set" methods. The idea is the following: if one knew in advance which constraints were active, it would be possible to cancel all of the inactive constraints which simplifies the problem.

The simplest method in this direction is known as *chunking*. It starts with an arbitrary subset ("chunk" = working set) of the data and trains the SVM using an optimizer on this subset. The algorithm then keeps the support vectors and deletes the others. Next, the M points (M algorithm parameter) from the remaining part of the data, where the "current SVM" makes the largest errors, are added to these support vectors to form a new chunk. This procedure is iterated. In general, the working set grows until in the last iteration the machine is trained on the set of support vectors representing the active constraints. Chunking techniques in SVMs were already used by Vapnik [106] and were improved and generalized in various papers.

Currently, more advanced "working set" methods, namely, *decomposition algorithms*, are one of the major tools to train SVMs. These methods select in each iteration a small *fixed size* subset of variables as working set, and a QP problem is solved with respect to this set (see, e.g., Osuna et al. [77]). A special type of decomposition methods is the *sequential minimal optimization* (SMO) which uses only *working sets of two variables*. This method was introduced by Platt [80] for classification; see Flake and Lawrence [42] for regression. The main advantage of these extreme small working sets is that the partial QP problems can be solved analytically. For the soft margin SVM in the dual form from Sect. 3 (with a variable exchange $\alpha \mapsto \mathbf{Y}\alpha$)

$$\frac{1}{2}\alpha^{\mathrm{T}}\mathbf{K}\alpha - \langle y, \alpha\rangle \quad \text{subject to} \quad \langle 1_m, \alpha\rangle = 0,\ 0 \le y\alpha \le C.$$

the SMO algorithm looks as follows:

SMO-type decomposition methods

1. Fix $\alpha^{(1)}$ as initial feasible solution and set $k := 1$.
2. If $\alpha^{(k)}$ solves the dual problem up to a desired precision, stop.
   Otherwise, select a working set $B := \{i, j\} \subset \{1, \ldots, m\}$. Define $N := \{1, \ldots, m\} \backslash B$
   and $\alpha_B^{(k)}$ and $\alpha_N^{(k)}$ as sub-vectors of $\alpha^{(k)}$ corresponding to B and N, resp.
3. Solve the following subproblem with fixed $\alpha_N^{(k)}$ for α_B:

$$\tfrac{1}{2}\left(\alpha_B^{\mathrm{T}}\,(\alpha_N^{(k)})^{\mathrm{T}}\right)\begin{pmatrix} K_{BB} & K_{BN} \\ K_{NB} & K_{NN}\end{pmatrix}\begin{pmatrix}\alpha_B \\ \alpha_N^{(k)}\end{pmatrix} - \left(y_B^{\mathrm{T}}\, y_N^{\mathrm{T}}\right)\begin{pmatrix}\alpha_B \\ \alpha_N^{(k)}\end{pmatrix}$$
$$= \tfrac{1}{2}\left(\alpha_i\ \alpha_j\right)\begin{pmatrix} K_{ii} & K_{ij} \\ K_{ij} & K_{jj}\end{pmatrix}\begin{pmatrix}\alpha_i \\ \alpha_j\end{pmatrix} - \left(\alpha_i\ \alpha_j\right)\left(K_{BN}\alpha_N^{(k)} - y_B\right) + \text{constant} \ \rightarrow \min_{\alpha_B}$$

   subject to $\alpha_i + \alpha_j = -1_{m-2}^{\mathrm{T}}\alpha_N^{(k)}, \quad 0 \le y_i\alpha_i, y_j\alpha_j \le C$.

   Set $\alpha_B^{(k+1)}$ to be the minimizer.
4. Set $\alpha_N^{(k+1)} := \alpha_N^{(k)}$, $k \mapsto k+1$ and goto Step 2.

The analytical solution in Step 3 is given as follows: for simplicity, set $\beta := -1_{m-2}^{\mathrm{T}}\alpha_N^{(k)}$ and $(c_i\, c_j)^{\mathrm{T}} := K_{BN}\alpha_N^{(k)} - y_B$. Substituting $\alpha_j = \beta - \alpha_i$ from the first constraint into the objective function, one gets

$$\frac{1}{2}\alpha_i^2(K_{ii} - 2K_{ij} + K_{jj}) + \alpha_i(\beta K_{ij} - \beta K_{jj} - c_i + c_j) + \text{constant} \ \rightarrow \min_{\alpha_i}.$$

If $\mathbf{K}$ is positive definite, it holds that $K_{ii} - 2K_{ij} + K_{jj} > 0$, and the above function has a unique finite global minimizer $\alpha_{i,g}$. One has to take care about the second constraint. This constraint requires that $\alpha_i \in [L, U]$, where L and U are defined by

$$(L, U) := \begin{cases} (\max(0, \beta - C), \min(C, \beta)) & \text{if } y_i = 1, y_j = 1, \\ (\max(0, \beta), \min(C, \beta + C)) & \text{if } y_i = 1, y_j = -1, \\ (\max(-C, \beta - C), \min(0, \beta)) & \text{if } y_i = -1, y_j = 1, \\ (\max(-C, \beta), \min(0, \beta + C)) & \text{if } y_i = -1, y_j = -1. \end{cases}$$

Hence the minimizer in Step 3 is given by $(\alpha_i^*, \beta - \alpha_i^*)$, where

$$\alpha_i := \begin{cases} \alpha_{ig} & \text{if } \alpha_{ig} \in [L, U], \\ L & \text{if } \alpha_{ig} < L, \\ U & \text{if } \alpha_{ig} > U. \end{cases}$$

It remains to determine the *selection of the working set*. (The determination of the stopping criteria is beyond the scope of this paper.) Indeed, current decomposition methods vary mainly according to different working set selections. The SVM$^{\text{light}}$ algorithm of Joachims [52], was originally based on a rule for the selection the working set of Zoutendijk [120]. Moreover, this algorithm uses a *shrinking* technique to speed up the computational time. Shrinking is based on the idea that if a variable $\alpha_i^{(k)}$ remains equal to zero or C for many iteration steps, then it will probably not change anymore. The variable can be removed from the optimization problem such that a more efficient overall optimization is obtained. Another shrinking implementation is used in the software package LIBSVM of Chang and Lin [22]. A modification of Joachims' algorithm for regression called "SVMTorch" was given by Collobert and Bengio [25]. An often addressed working set selection due to Keerthi et al. [55] is the so-called *maximal violating pair* strategy. A more general way of choosing the two-element working set, namely, by choosing a *constant factor violating pair*, was given, including a convergence proof, by Chen et al. [24]. For convergence results, see also the paper of Lin [65]. The *maximal violating pair* strategy relies on first-order (i.e., gradient) information of the objective function. Now for QP, second-order information directly relates to the decrease of the objective function. The paper of Fan et al. [38] proposes a promising working set selection based on second-order information.

For an overview of SVM solvers for large data sets, the reader may also consult the books of Huang et al. [50] and Bottou et al. [13] and the paper of Mangasagian and Musicant [69] with the references therein. An extensive list of SVM software including logistic loss functions and least squares loss functions can be found on the webpages www.kernel-machines.org and www.support-vector-machines.org.

6 Conclusion

The invention of SVMs in the 1990s led to an explosion of applications and theoretical results. This paper can only give a very basic introduction into the meanwhile classical techniques in this field. It is restricted to supervised learning although SVMs have also a large impact on semi- and unsupervised learning.

Some new developments are sketched as *multitask learning* where, in contrast to single-task learning, only limited work was involved until now and novel techniques taken from convex analysis come into the play.

An issue that is not addressed in this paper is the *robustness* of SVMs. There is some ongoing research on connections between stability, learning, and prediction of ERM methods (see, e.g., the papers of Elisseeff et al. [36] and Mukherjee et al. [74]).

Another field that has recently attained attention is the use of kernels as *diffusion kernels* on graphs (see Kondor and J. Lafferty [58] and also the book of Shawe-Taylor and Cristianini [88]).

Cross-References

- ▸ Duality and Convex Programming
- ▸ Large-Scale Inverse Problems in Imaging

References

1. Aizerman, M., Braverman, E., Rozonoer, L.: Uncovering shared structures in multiclass classification. In: International Conference on Machine Learning, pp. 821–837 (1964)
2. Amit, Y., Fink, M., Srebro, N., Ullman, S.: Theoretocal foundations of the potential function method in pattern recognition learning. Autom. Remote Control **25**, 17–24 (2007)
3. Anthony, M., Bartlett, P.L.: Neural Network Learning: Theoretical Foundations. Cambridge University Press, Cambridge (1999)
4. Argyriou, A., Evgeniou, T., Pontil, M.: Convex multi-task feature learning. Mach. Learn. **73**(3), 243–272 (2008)
5. Aronszajn, N.: Theory of reproducing kernels. Trans. Am. Math. Soc. **68**, 337–404 (1950)
6. Bartlett, P.L., Jordan, M.I., McAuliffe, J.D.: Convexity, classification, and risk bounds. J. Am. Stat. Assoc. **101**, 138–156 (2006)
7. Bennett, K.P., Mangasarian, O.L.: Robust linear programming discrimination of two linearly inseparable sets. Optim. Methods Softw. **1**, 23–34 (1992)
8. Berlinet, A., Thomas-Agnan, C.: Reproducing Kernel Hilbert Spaces in Probability and Statistics. Kluwer, Dordrecht (2004)
9. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer, New York (2006)
10. Björck, A.: Least Squares Problems. SIAM, Philadelphia (1996)
11. Bonnans, J.F., Shapiro, A.: Perturbation Analysis of Optimization Problems. Springer, New York (2000)
12. Boser, G.E., Guyon, I., Vapnik, V.: A training algorithm for optimal margin classifiers. In: Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory, Madison, pp. 144–152 (1992)
13. Bottou, L., Chapelle, L., DeCoste, O., Weston, J. (eds.): Large Scale Kernel Machines. MIT, Cambridge (2007)
14. Boucheron, S., Bousquet, O., Lugosi, G.: Theory of classification: a survey on some recent advances. ESAIM Probab. Stat. **9**, 323–375 (2005)
15. Bousquet, O., Elisseeff, A.: Algorithmic stability and generalization performance. In: Leen, T.K., Dietterich, T.G., Tresp, V. (eds.) Advances in Neural Information Processing Systems 13, pp. 196–202. MIT, Cambridge (2001)
16. Bradley, P.S., Mangasarian, O.L.: Feature selection via concave minimization and support vector machines. In: Proceedings of the 15th International Conference on Machine Learning, Madison, pp. 82–90. Morgan Kaufmann, San Francisco (1998)
17. Brown, M., Grundy, W., Lin, D., Cristianini, N., Sugnet, C., Furey, T., Ares, M., Haussler, D.: Knowledge-based analysis of microarray gene-expression data by using support vector machines. Proc. Natl. Acad. Sci. **97**(1), 262–267 (2000)
18. Buhmann, M.D.: Radial Basis Functions. Cambridge University Press, Cambridge (2003)
19. Burges, C.J.C.: A tutorial on support vector machines for pattern recognition. Data Min. Knowl. Discov. **2**(2), 121–167 (1998)
20. Cai, J.-F., Candès, E.J., Shen, Z.: A singular value thresholding algorithm for matrix completion. Technical report, UCLA Computational and Applied Mathematics (2008)
21. Caruana, R.: Multitask learning. Mach. Learn. **28**(1), 41–75 (1997)
22. Chang, C.-C., Lin, C.-J.: LIBSVM: a library for support vector machines. www.csie.ntu.edu.tw/~cjlin/papers/libsvm.ps.gz (2004)

23. Chapelle, O., Haffner, P., Vapnik, V.N.: SVMs for histogram-based image classification. IEEE Trans. Neural Netw. **10**(5), 1055–1064 (1999)
24. Chen, P.-H., Fan, R.-E., Lin, C.-J.: A study on SMO-type decomposition methods for support vector machines. IEEE Trans. Neural Netw. **17**, 893–908 (2006)
25. Collobert, R., Bengio, S.: Support vector machines for large scale regression problems. J. Mach. Learn. Res. **1**, 143–160 (2001)
26. Cortes, C., Vapnik, V.: Support vector networks. Mach. Learn. **20**, 273–297 (1995)
27. Cristianini, N., Shawe-Taylor, J.: An Introduction to Support Vector Machines. Cambridge University Press, Cambridge (2000)
28. Cucker, F., Smale, S.: On the mathematical foundations of learning. Bull. Am. Math. Soc. **39**, 1–49 (2002)
29. Cucker, F., Zhou, D.X.: Learning Theory: An Approximation Point of View. Cambridge University Press, Cambridge (2007)
30. Devroye, L., Györfi, L., Lugosi, G.: A Probabilistic Theory of Pattern Recognition. Springer, New York (1996)
31. Devroye, L.P.: Any discrimination rule can have an arbitrarily bad probability of error for finite sample size. IEEE Trans. Pattern Anal. Mach. Intell. **4**, 154–157 (1982)
32. Dietterich, T.G., Bakiri, G.: Solving multiclass learning problems via error-correcting output codes. J. Artif. Intell. Res. **2**, 263–286 (1995)
33. Dinuzzo, F., Neve, M., Nicolao, G.D., Gianazza, U.P.: On the representer theorem and equivalent degrees of freedom of SVR. J. Mach. Learn. Res. **8**, 2467–2495 (2007)
34. Duda, R.O., Hart, P.E., Stork, D.: Pattern Classification, 2nd edn. Wiley, New York (2001)
35. Edmunds, D.E., Triebel, H.: Function Spaces, Entropy Numbers, Differential Operators. Cambridge University Press, Cambridge (1996)
36. Elisseeff, A., Evgeniou, A., Pontil, M.: Stability of randomised learning algorithms. J. Mach. Learn. Res. **6**, 55–79 (2005)
37. Evgeniou, T., Pontil, M., Poggio, T.: Regularization networks and support vector machines. Adv. Comput. Math. **13**(1), 1–50 (2000)
38. Fan, R.-E., Chen, P.-H., Lin, C.-J.: Working set selection using second order information for training support vector machines. J. Mach. Learn. Res. **6**, 1889–1918 (2005)
39. Fasshauer, G.E.: Meshfree Approximation Methods with MATLAB. World Scientific, Hackensack (2007)
40. Fazel, M., Hindi, H., Boyd, S.P.: A rank minimization heuristic with application to minimum order system approximation. In: Proceedings of the American Control Conference, pp. 4734–4739 (2001)
41. Fisher, R.A.: The use of multiple measurements in taxonomic problems. Ann. Eugen. **7**, 179–188 (1936)
42. Flake, G.W., Lawrence, S.: Efficient SVM regression training with SMO. Technical report, NEC Research Institute (1999)
43. Gauss, C.F.: Theory of the Motion of the Heavenly Bodies Moving about the Sun in Conic Sections (C. H. Davis, Trans.) Dover, New York (1963). First published 1809
44. Girosi, F.: An equivalence between sparse approximation and support vector machines. Neural Comput. **10**(6), 1455–1480 (1998)
45. Golub, G.H., Loan, C.F.V.: Matrix Computation, 3rd edn. John Hopkins University Press, Baltimore (1996)
46. Györfi, L., Kohler, M., Krzyżak, A., Walk, H.: A Distribution-Free Theory of Nonparametric Regression. Springer, New York (2002)
47. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. Springer, New York (2001)
48. Herbrich, R.: Learning Kernel Classifiers: Theory and Algorithms. MIT, Cambridge (2001)
49. Hoerl, A.E., Kennard, R.W.: Ridge regression: biased estimation for nonorthogonal problems. Technometrics **12**(1), 55–67 (1970)
50. Huang, T., Kecman, V., Kopriva, I., Friedman, J.: Kernel Based Algorithms for Mining Huge Data Sets: Supervised, Semi-supervised and Unsupervised Learning. Springer, Berlin (2006)

51. Jaakkola, T.S., Haussler, D.: Probabilistic kerbnel regression models. In: Proceedings of the 1999 Conference on Artificial Intelligence and Statistics, Fort Lauderdale (1999)
52. Joachims, T.: Making large-scale SVM learning practical. In: Schölkopf, B., Burges, C., Smola, A. (eds.) Advances in Kernel Methods-Support Vector Learning, pp. 41–56. MIT, Cambridge (1999)
53. Joachims, T.: Learning to Classify Text Using Support Vector Machines. Kluwer Academic, Boston (2002)
54. Kailath, T.: RKHS approach to detection and estimation problems: part i: deterministic signals in Gaussian noise. IEEE Trans. Inf. Theory **17**(5), 530–549 (1971)
55. Keerthi, S.S., Shevade, S.K., Battacharyya, C., Murthy, K.R.K.: Improvements to Platt's SMO algorithm for SMV classifier design. Neural Comput. **13**, 637–649 (2001)
56. Kimeldorf, G.S., Wahba, G.: Some results on Tchebycheffian spline functions. J. Math. Anal. Appl. **33**, 82–95 (1971)
57. Kolmogorov, A.N., Tikhomirov, V.M.: ε-entropy and ε-capacity of sets in functional spaces. Am. Math. Soc. Transl. **17**, 277–364 (1961)
58. Kondor, R.I., Lafferty, J.: Diffusion kernels on graphs and other discrete structures. In: Kauffman, M. (ed.) Proceedings of the International Conference on Machine Learning (2002)
59. Krige, D.G.: A statistical approach to some basic mine valuation problems on the Witwatersrand. J. Chem. Metall. Min. Soc. S. Afr. **52**(6), 119–139 (1951)
60. Kuhn, H.W., Tucker, A.W.: Nonlinear programming. In: Proceedings Berkley Symposium on Mathematical Statistics and Probability, pp. 482–492. University of California Press (1951)
61. Laplace, P.S.: Théorie Analytique des Probabilités, 3rd edn. Courier, Paris (1816)
62. LeCun, Y., Jackel, L.D., Bottou, L., Brunot, A., Cortes, C., Denker, J.S., Drucker, H., Guyon, I., Müller, U., Säckinger, E., Simard, P., Vapnik, V.: Comparison of learning algorithms for handwritten digit recognition. In: Fogelman-Souleé, F., Gallinari, P. (eds.) Proceedings ICANN'95, Paris, vol. 2, pp. 53–60 (1995)
63. Legendre, A.M.: Nouvelles Méthodes pour la Determination des Orbites des Cométes. Courier, Paris (1805)
64. Leopold, E., Kinderman, J.: Text categogization with support vector machines how to represent text in input space? Mach. Learn. **46**(1–3), 223–244 (2002)
65. Lin, C.J.: On the convergence of the decomposition method for support vector machines. IEEE Trans. Neural Netw. **12**, 1288–1298 (2001)
66. Lu, Z., Monteiro, R.D.C., Yuan, M.: Convex optimization methods for dimension reduction and coefficient estimation in multivariate linear regression. Math. Program. **131**(1–2), 163–194 (2012)
67. Ma, S., Goldfarb, D., Chen, L.: Fixed point and Bregman iterative methods for matrix rank minimization. Technical report 08-78, UCLA Computational and Applied Mathematics (2008)
68. Mangasarian, O.L.: Nonlinear Programming. SIAM, Madison (1994)
69. Mangasarian, O.L., Musicant, D.R.: Successive overrelaxation for support vector machines. IEEE Trans. Neural Netw. **10**, 1032–1037 (1999)
70. Matheron, G.: Principles of geostatistics. Econ. Geol. **58**, 1246–1266 (1963)
71. Micchelli, C.A.: Interpolation of scattered data: distance matices and conditionally positive definite functions. Constr. Approx. **2**, 11–22 (1986)
72. Micchelli, C.A., Pontil, M.: On learning vector-valued functions. Neural Comput. **17**, 177–204 (2005)
73. Mitchell, T.M.: Machine Learning. McGraw-Hill, Boston (1997)
74. Mukherjee, S., Niyogi, P., Poggio, T., Rifkin, R.: Learning theory: stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. Adv. Comput. Math. **25**, 161–193 (2006)
75. Neumann, J., Schnörr, C., Steidl, G.: Efficient wavelet adaptation for hybrid wavelet–large margin classifiers. Pattern Recognit. **38**, 1815–1830 (2005)
76. Obozinski, G., Taskar, B., Jordan, M.I.: Joint covariate selection and joint subspace selection for multiple classification problems. Stat. Comput. **20**, 231–252 (2010)

77. Osuna, E., Freund, R., Girosi, F.: Training of support vector machines: an application to face detection. In: Proceedings of the CVPR'97, San Juan, pp. 130–136. IEEE Computer Society, Washington, DC (1997)
78. Parzen, E.: Statistical inference on time series by RKHS methods. Technical report, Department of Statistics, Stanford University (1970)
79. Pinkus, A.: N-width in Approximation Theory. Springer, Berlin/Heidelberg/New York-Tokyo (1996)
80. Platt, J.C.: Fast training of support vector machines using sequential minimal optimization. In: Schölkopf, B., Burges, C.J.C., Smola, A.J. (eds.) Advances in Kernel Methods – Support Vector Learning, pp. 185–208. MIT, Cambridge (1999)
81. Poggio, T., Girosi, F.: Networks for approximation and learning. Proc. IEEE **78**(9), 1481–1497 (1990)
82. Pong, T.K., Tseng, P., Ji, S., Ye, J.: Trace norm regularization: reformulations, algorithms and multi-task learning. SIAM J. Optim. **20**, 3465–3489 (2010)
83. Povzner, A.Y.: A class of Hilbert function spaces. Dokl. Akad. Nauk USSR **68**, 817–820 (1950)
84. Rosenblatt, F.: The perceptron: a probabilistic model for information storage and organization in the brain. Psychol. Rev. **65**, 386–408 (1959)
85. Schoenberg, I.J.: Metric spaces and completely monotone functions. Ann. Math. **39**, 811–841 (1938)
86. Schölkopf, B., Herbrich, R., Smola, A.J.: A generalized representer theorem. In: Helmbold, D., Williamson, B. (eds.) Proceedings of the 14th Annual Conference on Computational Learning Theory, Amsterdam, pp. 416–426. Springer, New York (2001)
87. Schölkopf, B., Smola, A.J.: Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT, Cambridge (2002)
88. Shawe-Taylor, J., Cristianini, N.: Kernel Methods for Pattern Analysis, 4th edn. Cambridge University Press, Cambridge (2009)
89. Smola, A.J., Schölkopf, B., Müller, K.R.: The connection between regularization operators and support vector kernels. Neural Netw. **11**, 637–649 (1998)
90. Spellucci, P.: Numerische Verfahren der Nichtlinearen Optimierung. Birkhäuser, Basel/Boston/Berlin (1993)
91. Srebro, N., Rennie, J.D.M., Jaakkola, T.S.: Maximum-margin matrix factorization. In: NIPS, pp. 1329–1336 (2005)
92. Steinwart, I.: Sparseness of support vector machines. J. Mach. Learn. Res. **4**, 1071–1105 (2003)
93. Steinwart, I., Christmann, A.: Support Vector Machines. Springer, New York (2008)
94. Stone, C.: Consistent nonparametric regression. Ann. Stat. **5**, 595–645 (1977)
95. Strauss, D.J., Steidl, G.: Hybrid wavelet-support vector classification of waveforms. J. Comput. Appl. Math. **148**, 375–400 (2002)
96. Strauss, D.J., Steidl, G., Delb, D.: Feature extraction by shape-adapted local discriminant bases. Signal Process. **83**, 359–376 (2003)
97. Sutton, R.S., Barton, A.G.: Reinforcement Learning: An Introduction. MIT, Cambridge (1998)
98. Suykens, J.A.K., Gestel, T.V., Brabanter, J.D., Moor, B.D., Vandewalle, J.: Least Squares Support Vector Machines. World Scientific, Singapore (2002)
99. Suykens, J.A.K., Vandevalle, J.: Least squares support vector machine classifiers. Neural Process. Lett. **9**(3), 293–300 (1999)
100. Tao, P.D., An, L.T.H.: A D.C. optimization algorithm for solving the trust-region subproblem. SIAM J. Optim. **8**(2), 476–505 (1998)
101. Tibshirani, R.: Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B **58**(1), 267–288 (1996)
102. Tikhonov, A.N., Arsenin, V.Y.: Solution of Ill-Posed Problems. Winston, Washington, DC (1977)

103. Toh, K.-C., Yun, S.: An accelerated proximal gradient algorithm for nuclear norm regularized least squares problems. Technical report, Department of Mathematics, National University of Singapore, Singapore (2009)
104. Tsypkin, Y.: Adaptation and Learning in Automatic Systems. Academic, New York (1971)
105. Vapnik, V.: Statistical Learning Theory. Wiley, New York (1998)
106. Vapnik, V.N.: Estimation of Dependicies Based on Empirical Data. Springer, New York (1982)
107. Vapnik, V.N., Chervonenkis, A.: Theory of Pattern Recognition (in Russian). Nauka, Moscow (1974) (German translation: Theorie der Zeichenerkennung, Akademie-Verlag, Berlin, 1979 edition)
108. Vapnik, V.N., Lerner, A.: Pattern recognition using generalized portrait method. Autom. Remote Control **24**, 774–780 (1963)
109. Vidyasagar, M.: A Theory of Learning and Generalization: With Applications to Neural Networks and Control Systems, 2nd edn. Springer, London (2002)
110. Viola, P., Jones, M.: Robust real-time face detection. Int. J. Comput. Vis. **57**(2), 137–154 (2004)
111. Vito, E.D., Rosasco, L., Caponnetto, A., Piana, M., Verri, A.: Some properties of regularized kernel methods. J. Mach. Learn. Res. **5**, 1363–1390 (2004)
112. Wahba, G.: Spline Models for Observational Data. SIAM, New York (1990)
113. Weimer, M., Karatzoglou, A., Smola, A.: Improving maximum margin matrix factorization. Mach. Learn. **72**(3), 263–276 (2008)
114. Wendland, H.: Scattered Data Approximation. Cambridge University Press, Cambridge (2005)
115. Weston, J., Elisseeff, A., Schölkopf, B., Tipping, M.: Use of the zero-norm with linear models and kernel methods. J. Mach. Learn. Res. **3**, 1439–1461 (2003)
116. Weston, J., Watkins, C.: Multi-class support vector machines. In: Verlysen, M. (ed.) Proceedings ESANN'99, Brussels. D-Facto Publications (1999)
117. Wolfe, P.: Duality theorem for nonlinear programming. Q. Appl. Math. **19**, 239–244 (1961)
118. Zdenek, D.: Optimal Quadratic Programming Algorithms with Applications to Variational Inequalities. Springer, New York (2009)
119. Zhang, T.: Statistical behaviour and consistency of classification methods based on convex risk minimization. Ann. Stat. **32**, 56–134 (2004)
120. Zoutendijk, G.: Methods of Feasible Directions. A Study in Linear and Nonlinear Programming. Elsevier, Amsterdam (1960)

Total Variation in Imaging

V. Caselles, A. Chambolle, and M. Novaga

Contents

1 Introduction 1456
2 Notation and Preliminaries on BV Functions 1460
Definition and Basic Properties 1460
Sets of Finite Perimeter: The Co-area Formula 1461
The Structure of the Derivative of a BV Function 1461
3 The Regularity of Solutions of the TV Denoising Problem 1463
The Discontinuities of Solutions of the TV Denoising Problem 1463
Hölder Regularity Results 1467
4 Some Explicit Solutions 1468
5 Numerical Algorithms: Iterative Methods 1471
Notation 1471
Chambolle's Algorithm 1473
Primal-Dual Approaches 1475
6 Numerical Algorithms: Maximum-Flow Methods 1477
Discrete Perimeters and Discrete Total Variation 1477
Graph Representation of Energies for Binary MRF 1479
7 Other Problems: Anisotropic Total Variation Models 1482
Global Solutions of Geometric Problems 1482
A Convex Formulation of Continuous Multilabel Problems 1486
8 Other Problems: Image Restoration 1488

V. Caselles (✉)
DTIC, Universitat Pompeu-Fabra, Barcelona, Spain
e-mail: vicent.caselles@upf.edu

A. Chambolle
CNRS UMR 7641, Ecole Polytechnique, Palaiseau Cedex, France
e-mail: antonin.chambolle@polytechnique.fr

M. Novaga
Dipartimento di Matematica, Università di Padova, Padova, Italy
e-mail: novaga@math.unipd.itt

© Springer Science+Business Media New York 2015

O. Scherzer (ed.), *Handbook of Mathematical Methods in Imaging*,
DOI 10.1007/978-1-4939-0790-8_23

Some Restoration Experiments 1491
The Image Model 1493
9 Final Remarks: A Different Total Variation-Based Approach to Denoising 1495
10 Conclusion 1496
References 1497

Abstract

The use of total variation as a regularization term in imaging problems was motivated by its ability to recover the image discontinuities. This is on the basis of his numerous applications to denoising, optical flow, stereo imaging and 3D surface reconstruction, segmentation, or interpolation, to mention some of them. On one hand, we review here the main theoretical arguments that have been given to support this idea. On the other hand, we review the main numerical approaches to solve different models where total variation appears. We describe both the main iterative schemes and the global optimization methods based on the use of max-flow algorithms. Then we review the use of anisotropic total variation models to solve different geometric problems and its use in finding a convex formulation of some non-convex total variation problems. Finally we study the total variation formulation of image restoration.

1 Introduction

The total variation model in image processing was introduced in the context of image restoration [55] and image segmentation, related to the study of the Mumford–Shah segmentation functional [34]. Being more related to our purposes here, let us consider the case of image denoising and restoration.

We assume that the degradation of the image occurs during image acquisition and can be modeled by a linear and translation invariant blur and additive noise:

$$f = h * u + n, \tag{1}$$

where $u : \mathbb{R}^2 \to \mathbb{R}$ denotes the ideal undistorted image, $h : \mathbb{R}^2 \to \mathbb{R}$ is a blurring kernel, f is the observed image which is represented as a function $f : \mathbb{R}^2 \to \mathbb{R}$, and n is an additive white noise with zero mean and standard deviation σ. In practice, the noise can be considered as Gaussian.

A particular and important case contained in the above formulation is the denoising problem which corresponds to the case where $h = \delta$, so that Eq. (1) is written as

$$f = u + n, \tag{2}$$

where n is an additive Gaussian white noise of zero mean and variance σ^2.

The problem of recovering u from f is ill-posed. Several methods have been proposed to recover u. Most of them can be classified as regularization methods

which may take into account statistical properties (Wiener filters), information theoretic properties [35], a priori geometric models [55], or the functional analytic behavior of the image given in terms of its wavelet coefficients (see [48] and references therein).

The typical strategy to solve this ill-conditioning is regularization [56]. In the linear case the solution of (1) is estimated by minimizing a functional

$$J_\gamma(u) = \| Hu - f \|_2^2 + \gamma \| Qu \|_2^2, \tag{3}$$

which yields the estimate

$$u_\gamma = (H^t H + \gamma Q^t Q)^{-1} H^t f, \tag{4}$$

where $Hu = h * u$, and Q is a regularization operator. Observe that to obtain u_γ we have to solve a system of linear equations. The role of Q is, on one hand, to move the small eigenvalues of H away from zero while leaving the large eigenvalues unchanged, and, on the other hand, to incorporate the a priori (smoothness) knowledge that we have on u.

If we treat u and n as random vectors and we select $\gamma = 1$ and $Q = R_s^{-1/2} R_n^{1/2}$ with R_s and R_n the image and noise covariance matrices, then (4) corresponds to the Wiener filter that minimizes the mean square error between the original and restored images.

One of the first regularization methods consisted in choosing between all possible solutions of (1) the one which minimized the Sobolev (semi) norm of u

$$\int_{\mathbb{R}^2} |Du|^2 \, dx, \tag{5}$$

which corresponds to the case $Qu = Du$. In the Fourier domain the solution of (3) given by (4) is $\hat{u} = \frac{\bar{\hat{h}}}{|\hat{h}|^2 + 4\gamma\pi^2|\xi|^2} \hat{f}$. From the above formula we see that high frequencies of f (hence, the noise) are attenuated by the smoothness constraint.

This formulation was an important step, but the results were not satisfactory, mainly due to the inability of the previous functional to resolve discontinuities (edges) and oscillatory textured patterns. The smoothness required by the finiteness of the Dirichlet integral (5) constraint is too restrictive. Indeed, functions in $W^{1,2}(\mathbb{R}^2)$ (i.e., functions $u \in L^2(\mathbb{R}^2)$ such that $Du \in L^2(\mathbb{R}^2)$) cannot have discontinuities along rectifiable curves. These observations motivated the introduction of Total Variation in image restoration problems by L. Rudin, S. Osher, and E. Fatemi in their work [55]. The a priori hypothesis is that functions of bounded variation (the BV model) [9] are a reasonable functional model for many problems in image processing, in particular, for restoration problems [55]. Typically, functions of bounded variation have discontinuities along rectifiable curves, being continuous in some sense (in the measure theoretic sense) away from discontinuities [9]. The discontinuities could be identified with edges. The ability of total variation regularization

to recover edges is one of the main features which advocates for the use of this model, but its ability to describe textures is less clear, even if some textures can be recovered, up to a certain scale of oscillation. An interesting experimental discussion of the adequacy of the BV-model to describe real images can be found in [41].

In order to work with images, we assume that they are defined in a bounded domain $\Omega \subseteq \mathbb{R}^2$ which we assume to be the interval $[0, N[^2$. As in most of the works, in order to simplify this problem, we shall assume that the functions h and u are periodic of period N in each direction. That amounts to neglecting some boundary effects. Therefore, we shall assume that h, u are functions defined in Ω and, to fix ideas, we assume that $h, u \in L^2(\Omega)$. Our problem is to recover as much as possible of u, from our knowledge of the blurring kernel h, the statistics of the noise n, and the observed image f.

On the basis of the BV model, Rudin–Osher–Fatemi [55] proposed to solve the following constrained minimization problem:

$$\begin{aligned} &\text{Minimize} \quad \int_\Omega |Du| \\ &\text{subject to} \quad \int_\Omega |h * u(x) - f(x)|^2 \, dx \leq \sigma^2 |\Omega|. \end{aligned} \tag{6}$$

Notice that the image acquisition model (1) is only incorporated through a global constraint. Assuming that $h * 1 = 1$ (energy preservation), the additional constraint that $\int_\Omega h * u \, dx = \int_\Omega f(x)$ is automatically satisfied by its minima [28]. In practice, the above problem is solved via the following unconstrained minimization problem:

$$\text{Minimize} \int_\Omega |Du| + \frac{1}{2\lambda} \int_\Omega |h * u - f|^2 \, dx \tag{7}$$

where the parameter λ is positive. Recall that we may interpret λ as a penalization parameter which controls the trade-off between the goodness of fit of the constraint and the smoothness term given by the Total Variation. In this formulation, a methodology is required for a correct choice of λ. The connections between (6) and (7) were studied by A. Chambolle and P.L. Lions in [28] where they proved that both problems are equivalent for some positive value of the Lagrange multiplier λ.

In the denoising case, the unconstrained variational formulation (7) with $h = \delta$ is

$$\text{Minimize} \int_\Omega |Du| + \frac{1}{2\lambda} \int_\Omega |u - f|^2 \, dx, \tag{8}$$

and it has been the object of much theoretical and numerical research (see [10, 56] for a survey). Even if this model represented a theoretical and practical progress in the denoising problem due to the introduction of BV functions as image models, the experimental analysis readily showed its main drawbacks. Between them, let us mention the staircasing effect (when denoising a smooth ramp plus noise, the staircase is an admissible result), the pixelization of the image at smooth regions,

and the loss of fine textured regions, to mention some of them. This can be summarized with the simple observation that the residuals $f-u$, where u represents the solution of (8), do not look like noise. This has motivated the development of nonlocal filters [17] for denoising, the use of a stochastic optimization technique to estimate u [47], or the consideration of the image acquisition model as a set of local constraints [3, 38] to be discussed below.

Let us finally mention that, following the analysis of Y. Meyer in [48], the solution u of (8) permits to obtain a decomposition of the data f as a sum of two components $u+v$ where u contains the geometric sketch of f while v is supposed to contain its noise and textured parts.As Meyer observed, the L^2 norm of the residual $v := f - u$ in (8) is not the right one to obtain a decomposition of f in terms of geometry plus texture and he proposed to measure the size of the textured part v in terms of a dual BV norm showing that some models of texture have indeed a small dual BV norm.

In spite of its limitations, the total variation model has become one of the basic image models and has been adapted to many tasks: optical flow, stereo imaging and 3D surface reconstruction, segmentation, interpolation, or the study of $u + v$ models to mention a few cases. On the other hand, when compared to other robust regularization terms, it combines simplicity and geometric character and makes it possible a rigorous analysis. The theoretical analysis of the behavior of solutions of (8) has been the object of several works [6, 13, 14, 20, 48, 50] and will be summarized in Sects. 3 and 4.

Recall that one of the main reasons to introduce the Total Variation as a regularization term in imaging problems was its ability to recover discontinuities in the solution. This intuition has been confirmed by the experimental evidence and has been the motivation for the study of the local regularity properties of (8) in [20, 23]. After recalling in Sect. 2 some basic notions and results in the theory of bounded variation functions, we prove in section "The Discontinuities of Solutions of the TV Denoising Problem" that the set of jumps (in the BV sense) of the solution of (8) is contained in the set of jumps of the datum f [20]. In other words, model (8) does not create any new discontinuity besides the existing ones. As a refinement of the above statement, the local Hölder regularity of the solutions of (8) is studied in section "Hölder Regularity Results." This has to be combined with results describing which discontinuities are preserved. No general statement in this sense exists, but many examples are described in the papers [5, 10, 13, 14]. The preservation of a jump discontinuity depends on the curvature of the level line at the given point, the size of the jump, and the regularization parameter λ. This is illustrated in the example given in Sect. 4. The examples support the idea that total variation is not perfect but may be a reasonable regularization term in order to restore discontinuities.

Being considered as a basic model, the numerical analysis of the total variation model has been the object of intensive research. Many numerical approaches have been proposed in order to give fast, efficient methods which are also versatile to cover the whole range of applications. In Sect. 5 we review some basic iterative methods introduced to solve the Euler–Lagrange equations of (8). In particular, we review in section "Chambolle's Algorithm" the dual approach introduced by A.

Chambolle in [25]. In section "Primal-Dual Approaches" we review the primal-dual scheme of Zhu and Chan [57]. Both of them are between the most popular schemes by now. In Sect. 6 we discuss global optimization methods based on graph-cut techniques adapted to solve a quantized version of (8). Those methods have also become very popular due to its efficiency and versatility in applications and are an active area of research, as it can be seen in the references. Then, in section "Global Solutions of Geometric Problems" we review the applications of anisotropic TV problems to find the global solution of geometric problems. Similar anisotropic TV formulations appear as convexifications of nonlinear energies for disparity computation in stereo imaging, or related problems [30, 52], and they are reviewed in section "A Convex Formulation of Continuous Multilabel Problems."

In Sect. 8 we review the application of Total Variation in image restoration (6), describing the approach where the image acquisition model is introduced as a set of local constraints [3, 38, 54].

We could not close this chapter without reviewing in Sect. 9 a recent algorithm introduced by C. Louchet and L. Moisan [47] which uses a Bayesian approach leading to an estimate of u as the expected value of the posterior distribution of u given the data f. This estimate requires to compute an integral in a high dimensional space and the authors use a Monte-Carlo method with Markov Chain (MCMC) [47]. In this context, the minimization of the discrete version of (8) corresponds to a Maximum a Posterior (MAP) estimate of u.

2 Notation and Preliminaries on BV Functions

Definition and Basic Properties

Let Ω be an open subset of $\mathbb{R}^N$. Let $u \in L^1_{\mathrm{loc}}(\Omega)$. Recall that the distributional gradient of u is defined by

$$\int_\Omega \sigma \cdot Du = -\int_\Omega u(x)\mathrm{div}\,\sigma(x)\,dx \;\forall C_c^\infty(\Omega, \mathbb{R}^N),$$

where $C_c^\infty(\Omega; \mathbb{R}^N)$ denotes the vector fields with values in $\mathbb{R}^N$ which are infinitely differentiable and have compact support in Ω. The total variation of u in Ω is defined by

$$V(u, \Omega) := \sup \left\{ \int_\Omega u \,\mathrm{div}\,\sigma \; dx : \sigma \in C_c^\infty(\Omega; \mathbb{R}^N), |\sigma(x)| \leq 1 \;\forall x \in \Omega \right\}, \qquad (9)$$

where for a vector $v = (v_1, \ldots, v_N) \in \mathbb{R}^N$ we set $|v|^2 := \sum_{i=1}^N v_i^2$. Following the usual notation, we will denote $V(u, \Omega)$ by $|Du|(\Omega)$ or by $\int_\Omega |Du|$.

Definition 1. Let $u \in L^1(\Omega)$. We say that u is a function of bounded variation in Ω if $V(u, \Omega) < \infty$. The vector space of functions of bounded variation in Ω will be denoted by $\mathrm{BV}(\Omega)$.

Using Riesz representation Theorem [9], the above definition can be rephrased by saying that u is a function of bounded variation in Ω if the gradient Du in the sense of distributions is a (vector valued) Radon measure with finite total variation $V(u, \Omega)$.

Recall that $\mathrm{BV}(\Omega)$ is a Banach space when endowed with the norm $\|u\| := \int_\Omega |u|\ dx + |Du|(\Omega)$. Recall also that the map $u \to |Du|(\Omega)$ is $L^1_{\mathrm{loc}}(\Omega)$-lower semicontinuous, as a sup (9) of continuous linear forms [9].

Sets of Finite Perimeter: The Co-area Formula

Definition 2. A measurable set $E \subseteq \Omega$ is said to be of finite perimeter in Ω if $\chi_E \in \mathrm{BV}(\Omega)$. The perimeter of E in Ω is defined as $P(E, \Omega) := |D\chi_E|(\Omega)$. If $\Omega = \mathbb{R}^N$, we denote the perimeter of E in $\mathbb{R}^N$ by $P(E)$.

The following inequality holds for any two sets $A, B \subseteq \Omega$:

$$P(A \cup B, \Omega) + P(A \cap B, \Omega) \leq P(A, \Omega) + P(B, \Omega). \tag{10}$$

Theorem 1. *Let $u \in \mathrm{BV}(\Omega)$. Then for a.e. $t \in \mathbb{R}$ the set $\{u > t\}$ is of finite perimeter in Ω and one has*

$$\int_\Omega |Du| = \int_{-\infty}^{\infty} P(\{u > t\}, \Omega)\, dt.$$

In other words, the total variation of u amounts to the sum of the perimeters of its upper level sets.

An analogous formula with the lower level sets is also true. For a proof we refer to [9].

The Structure of the Derivative of a BV Function

Let us denote by $\mathcal{L}^N$ and $\mathcal{H}^{N-1}$, respectively, the N-dimensional Lebesgue measure and the $(N-1)$-dimensional Hausdorff measure in $\mathbb{R}^N$ (see [9] for precise definitions).

Let $u \in [L^1_{\mathrm{loc}}(\Omega)]^m$ $(m \geq 1)$. We say that u has an approximate limit at $x \in \Omega$ if there exists $\xi \in \mathbb{R}^m$ such that

$$\lim_{\rho \downarrow 0} \frac{1}{|B(x, \rho)|} \int_{B(x,\rho)} |u(y) - \xi| dy = 0. \tag{11}$$

The set of points where this does not hold is called the approximate discontinuity set of u, and is denoted by S_u. Using Lebesgue's differentiation theorem, one can show that the approximate limit ξ exists at $\mathcal{L}^N$-a.e. $x \in \Omega$, and is equal to $u(x)$, in particular, $|S_u| = 0$. If $x \in \Omega \setminus S_u$, the vector ξ is uniquely determined by (11) and we denote it by $\tilde{u}(x)$.

We say that u is approximately continuous at x if $x \notin S_u$ and $\tilde{u}(x) = u(x)$, that is, if x is a Lebesgue point of u (with respect to the Lebesgue measure).

Let $u \in [L^1_{\mathrm{loc}}(\Omega)]^m$ and $x \in \Omega \setminus S_u$; we say that u is approximately differentiable at x if there exists an $m \times N$ matrix L such that

$$\lim_{\rho \downarrow 0} \frac{1}{|B(x,\rho)|} \int_{B(x,\rho)} \frac{|u(y) - \tilde{u}(x) - L(y-x)|}{\rho}\, dy \;=\; 0. \tag{12}$$

In that case, the matrix L is uniquely determined by (12) and is called the approximate differential of u at x.

For $u \in \mathrm{BV}(\Omega)$, the gradient Du is a N-dimensional Radon measure that decomposes into its absolutely continuous and singular parts $Du = D^a u + D^s u$. Then $D^a u = \nabla u\, dx$ where ∇u is the Radon–Nikodym derivative of the measure Du with respect to the Lebesgue measure in $\mathbb{R}^N$. The function u is approximately differentiable $\mathcal{L}^N$-a.e. in Ω and the approximate differential coincides with $\nabla u(x)$ $\mathcal{L}^N$-a.e. The singular part $D^s u$ can be also split into two parts: the *jump* part $D^j u$ and the *Cantor* part $D^c u$.

We say that $x \in \Omega$ is an *approximate jump point of* u if there exist $u^+(x) \neq u^-(x) \in \mathbb{R}$ and $|\nu_u(x)| = 1$ such that

$$\lim_{\rho \downarrow 0} \frac{1}{|B^+_\rho(x, \nu_u(x))|} \int_{B^+_\rho(x,\nu_u(x))} |u(y) - u^+(x)|\, dy = 0$$

$$\lim_{\rho \downarrow 0} \frac{1}{|B^-_\rho(x, \nu_u(x))|} \int_{B^-_\rho(x,\nu_u(x))} |u(y) - u^-(x)|\, dy = 0,$$

where $B^+_\rho(x,\nu_u(x)) = \{y \in B(x,\rho) \,:\, \langle y - x, \nu_u(x)\rangle > 0\}$ and $B^-_\rho(x,\nu_u(x)) = \{y \in B(x,\rho) \,:\, \langle y - x, \nu_u(x)\rangle < 0\}$. We denote by J_u the set of approximate jump points of u. If $u \in \mathrm{BV}(\Omega)$, the set S_u is countably $\mathcal{H}^{N-1}$ rectifiable, J_u is a Borel subset of S_u, and $\mathcal{H}^{N-1}(S_u \setminus J_u) = 0$ [9]. In particular, we have that $\mathcal{H}^{N-1}$-a.e. $x \in \Omega$ is either a point of approximate continuity of $\tilde{u}$ or a jump point with two limits in the above sense. Finally, we have

$$D^j u = D^s u\llcorner_{J_u} = (u^+ - u^-)\nu_u \mathcal{H}^{N-1}\llcorner_{J_u} \quad \text{and} \quad D^c u = D^s u\llcorner_{(\Omega \setminus S_u)}.$$

For a comprehensive treatment of functions of bounded variation we refer to [9].

3 The Regularity of Solutions of the TV Denoising Problem

The Discontinuities of Solutions of the TV Denoising Problem

Given a function $f \in L^2(\Omega)$ and $\lambda > 0$ we consider the minimum problem

$$\min_{u \in \mathrm{BV}(\Omega)} \int_\Omega |Du| + \frac{1}{2\lambda} \int_\Omega (u - f)^2 \, dx \,. \tag{13}$$

Notice that problem (13) always admits a unique solution u_λ, since the energy functional is strictly convex.

As we mentioned in Sect. 1, one of the main reasons to introduce the Total Variation as a regularization term in imaging problems was its ability to recover the discontinuities in the solution. This section together with section "Hölder Regularity Results" and Sect. 4 is devoted to analyze this assertion. In this section we prove that the set of jumps of u_λ (in the BV sense) is contained in the set of jumps of f, whenever f has bounded variation. Thus, model (13) does not create any new discontinuity besides the existing ones. Section "Hölder Regularity Results" is devoted to review a local Hölder regularity result of [23]: the local Hölder regularity of the data is inherited by the solution. This has to be combined with results describing which discontinuities are preserved. In Sect. 4 we give an example of explicit solution of (13) which shows that the preservation of a jump discontinuity depends on the curvature of the level line at the given point, the size of the jump, and the regularization parameter λ. Other examples are given in the papers [2, 5, 10, 13, 14]. The examples support the idea that total variation may be a reasonable regularization term in order to restore discontinuities.

Let us recall the following observation, which is proved in [5, 18, 24].

Proposition 1. *Let u_λ be the (unique) solution of (13). Then, for any $t \in \mathbb{R}$, $\{u_\lambda > t\}$ (respectively, $\{u_\lambda \geq t\}$) is the minimal (resp., maximal) solution of the minimal surface problem*

$$\min_{E \subseteq \Omega} P(E, \Omega) + \frac{1}{\lambda} \int_E (t - f(x)) \, dx \tag{14}$$

(whose solution is defined in the class of finite-perimeter sets, hence up to a Lebesgue-negligible set). In particular, for all $t \in \mathbb{R}$ but a countable set, $\{u_\lambda = t\}$ has zero measure and the solution of (14) is unique up to a negligible set.

A proof that $\{u_\lambda > t\}$ and $\{u_\lambda \geq t\}$ both solve (14) is found in [24, Prop. 2.2]. A complete proof of this proposition, which we do not give here, follows from the co-area formula, which shows that, up to a renormalization, for any $u \in \mathrm{BV}(\Omega) \cap L^2(\Omega)$,

$$\int_\Omega |Du| + \frac{1}{2\lambda}\int_\Omega (u-f)^2\,dx = \int_{\mathbb{R}} \left(P\left(\{u>t\},\Omega\right) + \frac{1}{\lambda}\int_{\{u>t\}} (t-f)\,dx \right) dt\,,$$

and from the following comparison result for solutions of (14) which is proved in [5, Lemma 4]:

Lemma 1. *Let $f, g \in L^1(\Omega)$ and E and F be respectively minimizers of*

$$\min_E P(E,\Omega) - \int_E f(x)\,dx \quad \text{and} \quad \min_F P(F,\Omega) - \int_F g(x)\,dx\,.$$

Then, if $f < g$ a.e., $|E \setminus F| = 0$ (in other words, $E \subseteq F$ up to a negligible set).

Proof. Observe that we have

$$P(E,\Omega) - \int_E f(x)\,dx \le P(E\cap F,\Omega) - \int_{E\cap F} f(x)\,dx$$

$$P(F,\Omega) - \int_F g(x)\,dx \le P(E\cup F,\Omega) - \int_{E\cup F} g(x)\,dx.$$

Adding both inequalities and using that for two sets of finite perimeter we have (10) $P(E\cap F,\Omega) + P(E\cup F,\Omega) \le P(E,\Omega) + P(F,\Omega)$, we obtain that

$$\int_{E\setminus F} (g(x) - f(x))\,dx \le 0.$$

Since $g(x) - f(x) > 0$ a.e., this implies that $E \setminus F$ is a null set. □

The proof of this last lemma is easily generalized to other situations (Dirichlet boundary conditions, anisotropic and/or nonlocal perimeters; see [5] and also [2] for a similar general statement). Eventually, we mention that the result of Proposition 1 remains true if the term $(u(x) - f(x))^2/(2\lambda)$ in (13) is replaced with a term of the form $\Psi(x, u(x))$, with Ψ of class C^1 and strictly convex in the second variable, and replacing $(t - f(x))/\lambda$ with $\partial_u \Psi(x,t)$ in (14).

From Proposition 1 and the regularity theory for surfaces of prescribed curvature (see for instance [8]), we obtain the following regularity result (see also [2]).

Corollary 1. *Let $f \in L^p(\Omega)$, with $p > N$. Then, for all $t \in \mathbb{R}$ the super-level set $E_t := \{u_\lambda > t\}$ (respectively, $\{u_\lambda \ge t\}$) has boundary of class $C^{1,\alpha}$, for all $\alpha < (p-N)/p$, out of a closed singular set Σ of Hausdorff dimension at most $N-8$. Moreover, if $p = \infty$, the boundary of E_t is of class $W^{2,q}$ out of Σ, for all $q < \infty$, and is of class $C^{1,1}$ if $N = 2$.*

We now show that the jump set of u_λ is always contained in the jump set of f. Before stating this result let us recall two simple lemmas.

Lemma 2. *Let U be an open set in $\mathbb{R}^N$ and $v \in W^{2,p}(U)$, $p \geq 1$. We have that*

$$\operatorname{div}\left(\frac{\nabla v}{\sqrt{1+|\nabla v|^2}}\right)(y) = \operatorname{Trace}\left(A(\nabla v(y))D^2 v(y)\right) \qquad \text{a.e. in } U\ ,$$

where $A(\xi) = \frac{1}{(1+|\xi|^2)^{\frac{1}{2}}}\left(\delta_{ij} - \frac{\xi_i \xi_j}{(1+|\xi|^2)}\right)_{i,j=1}^{N}$, $\xi \in \mathbb{R}^N$.

The proof follows simply by taking $\varphi \in C_0^\infty(U)$, integrating by parts in U, and regularizing v with a smoothing kernel.

Lemma 3. *Let U be an open set in $\mathbb{R}^N$ and $v \in W^{2,1}(U)$. Assume that u has a minimum at $y_0 \in U$ and*

$$\lim_{\rho \to 0+} \frac{1}{B_\rho(y_0)} \int_{B_\rho(y_0)} \frac{|u(y) - u(y_0) - \nabla u(y_0)\cdot(y-y_0) - \frac{1}{2}\langle D^2 v(y_0)(y-y_0), y-y_0\rangle|}{\rho^2}\, dy = 0. \tag{15}$$

Then $D^2 v(y_0) \geq 0$.

If A is a symmetric matrix and we write $A \geq 0$ (respectively, $A \leq 0$) we mean that A is positive (resp., negative) semidefinite.

The result follows by proving that $\mathcal{H}^{N-1}$-a.e. for ξ in S^{N-1} (the unit sphere in $\mathbb{R}^N$) we have $\langle D^2 v(y_0)\xi, \xi\rangle \geq 0$.

Recall that if $v \in W^{2,1}(U)$, then (15) holds a.e. on U [58, Theorem 3.4.2].

Theorem 2. *Let $f \in \mathrm{BV}(\Omega) \cap L^\infty(\Omega)$. Then, for all $\lambda > 0$,*

$$J_{u_\lambda} \subseteq J_f \tag{16}$$

(up to a set of zero $\mathcal{H}^{N-1}$-measure).

Before giving the proof let us explain its main idea which is quite simple. Notice that, by (14), formally the Euler–Lagrange equation satisfied by ∂E_t is

$$\kappa_{E_t} + \frac{1}{\lambda}(t - f) = 0 \qquad \text{on } \partial E_t,$$

where κ_{E_t} is the sum of the principal curvatures at the points of ∂E_t. Thus if $x \in J_{u_\lambda} \setminus J_f$, then we may find two values $t_1 < t_2$ such that $x \in \partial E_{t_1} \cap \partial E_{t_2} \setminus J_f$. Notice that $E_{t_2} \subseteq E_{t_1}$ and the boundaries of both sets have a contact at x. Of the two, the smallest level set is the highest and has smaller mean curvature. This contradicts its contact at x.

Proof. Let us first recall some consequences of Corollary 1. Let $E_t := \{u_\lambda > t\}$, $t \in \mathbb{R}$, and let Σ_t be its singular set given by Corollary 1. Since $f \in L^\infty(\Omega)$, around each point $x \in \partial E_t \setminus \Sigma_t$, $t \in \mathbb{R}$, ∂E_t is locally the graph of a function in $W^{2,p}$ for all $p \in [1, \infty)$ (hence $C^{1,\alpha}$ for any $\alpha \in (0, 1)$). Moreover, if $\mathcal{N} := \bigcup_{t\in\mathbb{Q}} \Sigma_t$, then $\mathcal{H}^{N-1}(\mathcal{N}) = 0$.

Let us prove that $\mathcal{H}^{N-1}(J_{u_\lambda} \setminus J_f) = 0$. Observe that we may write [9]

$$J_{u_\lambda} = \bigcup_{t_1,t\in\mathbb{Q},t_1<t_2} \partial E_{t_1} \cap \partial E_{t_2}.$$

Thus it suffices to prove that for all $t_1, t_2 \in \mathbb{Q}$, $t_1 < t_2$, we have

$$\mathcal{H}^{N-1}\left(\partial E_{t_1} \cap \partial E_{t_2} \setminus \left(\mathcal{N} \cup J_f\right)\right) = 0. \tag{17}$$

Let us denote by B'_R the ball of radius $R > 0$ in $\mathbb{R}^{N-1}$ centered at 0. Let $C_R := B'_R \times (-R, R)$. Let us fix $t_1, t_2 \in \mathbb{Q}$, $t_1 < t_2$. Given $x \in \partial E_{t_1} \cap \partial E_{t_2} \setminus \mathcal{N}$, by Corollary 1, we know that there is some $R > 0$ such that, after a change of coordinates that aligns the x_N-axis with the normal to $\partial E_{t_1} \cap \partial E_{t_2}$ at x, we may write the set $\partial E_{t_i} \cap C_R$ as the graph of a function $v_i \in W^{2,p}(B'_R)$, $\forall p \in [1, \infty)$, $x = (0, v_i(0)) \in C_R \subseteq \Omega$, $\nabla v_i(0) = 0$, $i \in \{1, 2\}$. Without loss of generality, we assume that $v_i > 0$ in B'_R, and that E_{t_i} is the supergraph of v_i, $i = 1, 2$. From $t_1 < t_2$ and Lemma 1, it follows $E_{t_2} \subseteq E_{t_1}$, which gives in turn $v_2 \geq v_1$ in B'_R.

Notice that, since ∂E_{t_i} is of finite $\mathcal{H}^{N-1}$ measure, we may cover $\partial E_{t_1} \cap \partial E_{t_2} \setminus \mathcal{N}$ by a countable set of such cylinders. Thus, it suffices to prove that

$$\mathcal{H}^{N-1}\left((\partial E_{t_1} \cap \partial E_{t_2} \cap C_R) \setminus \left(\mathcal{N} \cup J_f\right)\right) = 0. \tag{18}$$

holds for any such cylinder C_R as constructed in the last paragraph.

Let us denote the points $x \in C_R$ as $x = (y, z) \in B'_R \times (-R, R)$. Then (18) will follow if we prove that

$$\mathcal{H}^{N-1}(\mathcal{M}_R) = 0, \tag{19}$$

where

$$\mathcal{M}_R := \{y \in B'_R : v_1(y) = v_2(y)\} \setminus \{y \in B'_R : (y, v_1(y)) \in J_f\}.$$

Recall that, by Theorem 3.108 in [9], $\mathcal{H}^{N-1}$-a.e. in $y \in B'_R$, the function $f(y, \cdot) \in \mathrm{BV}((-R, R))$ and the jumps of $f(y, \cdot)$ are the points z such that

$(y,z) \in J_f$. Recall that v_i is a local minimizer of

$$\min_v \mathcal{E}_i(v) := \int_{B'_R} \sqrt{1+|\nabla v|^2}\, dy - \frac{1}{\lambda} \int_{B'_R} \int_0^{v(y)} (t_i - f(y,z))\, dz\, dy.$$

By taking a positive smooth test function $\psi(y)$ of compact support in B'_R, and computing $\lim_{\epsilon \to 0+} \frac{1}{\epsilon} (\mathcal{E}_i(v+\epsilon\psi) - \mathcal{E}_i(v)) \geq 0$, we deduce that

$$\operatorname{div} \frac{\nabla v_i(y)}{\sqrt{1+|\nabla v_i(y)|^2}} + \frac{1}{\lambda} (t_i - f(y, v_i(y)+0)) \leq 0, \qquad \mathcal{H}^{N-1}\text{-a.e. in } B'_R. \tag{20}$$

In a similar way, we have

$$\operatorname{div} \frac{\nabla v_i(y)}{\sqrt{1+|\nabla v_i(y)|^2}} + \frac{1}{\lambda} (t_i - f(y, v_i(y)-0)) \geq 0, \qquad \mathcal{H}^{N-1}\text{-a.e. in } B'_R. \tag{21}$$

Finally we observe that since $v_1, v_2 \in W^{2,p}(B'_R)$ for any $p \in [1,\infty)$ and $v_2 \geq v_1$ in B'_R, by Lemma 3 we have that $D^2(v_1 - v_2)(y) \leq 0$ $\mathcal{H}^{N-1}$-a.e. on $\{y \in B'_R : v_1(y) = v_2(y)\}$.

Thus, if $\mathcal{H}^{N-1}(\mathcal{M}_R) > 0$, then there is a point $\bar{y} \in \mathcal{M}_R$ such that $\nabla v_1(\bar{y}) = \nabla v_2(\bar{y})$, $D^2(v_1 - v_2)(\bar{y}) \leq 0$, $f(\bar{y}, \cdot)$ is continuous at $v_1(\bar{y}) = v_2(\bar{y})$, and both Eqs. (20) and (21) hold at $\bar{y}$. As a consequence, using Lemma 2 and subtracting the two equations, we obtain

$$0 \geq \operatorname{trace}(A(\nabla v_1(\bar{y}))D^2 v_1(\bar{y})) - \operatorname{trace}(A(\nabla v_2(\bar{y}))D^2 v_2(\bar{y})) = \frac{t_2 - t_1}{\lambda} > 0,$$

This contradiction proves (19). □

Hölder Regularity Results

Let us review the local regularity result proved in [23]: if the datum f is locally Hölder continuous with exponent $\beta \in [0,1]$ in some region $\Omega' \subset \Omega$, then a local minimizer u of (13) is also locally Hölder continuous in Ω' with the same exponent.

Recall that a function $u \in \mathrm{BV}(\Omega)$ is a local minimizer of (13) if for any $v \in \mathrm{BV}(\Omega)$ such that $u - v$ has support in a compact subset $K \subset \Omega$, we have

$$\int_K |Du| + \frac{1}{2} \int_K |u(x) - f(x)|^2\, dx \;\leq\; \int_K |Dv| + \frac{1}{2} \int_K |v(x) - f(x)|^2\, dx \tag{22}$$

It follows that u satisfies the equation [18]

$$-\operatorname{div} z + u \;=\; f \tag{23}$$

with $z \in L^\infty(\Omega, \mathbb{R}^N)$ with $\|z\|_\infty \leq 1$, and $z \cdot Du = |Du|$ [10].

As in section "The Discontinuities of Solutions of the TV Denoising Problem" [20], the analysis of the regularity of the local minimizers of u is based on the following observation: for any $t \in \mathbb{R}$, the level sets $\{u > t\}$ (resp., $\{u \geq t\}$) are solutions (the minimal and maximal, indeed) of the prescribed curvature problem (14) which is defined in the class of finite-perimeter sets and hence up to a Lebesgue-negligible set. The local regularity of u can be described in terms of the distance of any two of its level sets. This is the main idea in [20] which can be refined to obtain the Hölder regularity of solutions of (14). As we argued in section "The Discontinuities of Solutions of the TV Denoising Problem," outside the jump discontinuities of f (modulo an $\mathcal{H}^{N-1}$-null set), any two level sets at different heights cannot touch and hence the function u is continuous there. To be able to assert a Hölder type regularity property for u one needs to prove a local estimate of the distance of the boundaries of two level sets. This can be done here under the assumption of local Hölder regularity for f [23].

Theorem 3. *Let $N \leq 7$ and let u be a solution of (23). Assume that f is in $C^{0,\beta}$ locally in some open set $A \subseteq \Omega$, for some $\beta \in [0,1]$. Then u is also $C^{0,\beta}$ locally in A.*

The Lipschitz case corresponds to $\beta = 1$.

One can also state a global regularity result for solutions of the Neumann problem when $\Omega \subset \mathbb{R}^N$ is a convex domain. Let $f : \overline{\Omega} \to \mathbb{R}$ be a uniformly continuous function, with modulus of continuity $\omega_f : [0, +\infty) \to [0, +\infty)$, that is, $|f(x) - f(y)| \leq \omega_f(|x - y|)$ for all $x, y \in \Omega$. We consider the solution u of (23) with homogeneous Neumann boundary condition, that is, such that (22) for any compact set $K \subset \overline{\Omega}$ and any $v \in \mathrm{BV}(\Omega)$ such that $v = u$ out of K. This solution is unique, as can be shown adapting the proof of [18, Cor. C.2.] (see also [10] for the required adaptations to deal with the boundary condition), which deals with the case $\Omega = \mathbb{R}^N$.

Then, the following result holds true [23]:

Theorem 4. *Assume $N \leq 7$. Then, the function u is uniformly continuous in Ω, with modulus $\omega_u \leq \omega_f$.*

Again, it is quite likely here that the assumption $N \leq 7$ is not necessary for this result.

4 Some Explicit Solutions

Recall that a convex body in $\mathbb{R}^N$ is a compact convex subset of $\mathbb{R}^N$. We say that a convex body is nontrivial if it has nonempty interior.

We want to exhibit the explicit solution of (13) when $f = \chi_C$ and C is a nontrivial convex body in $\mathbb{R}^N$. This will show that the preservation of a jump

discontinuity depends on the curvature of ∂C at the given point, the size of the jump, and the regularization parameter λ.

Let $u_{\lambda,C}$ be the unique solution of the problem:

$$\min_{u\in \mathrm{BV}(\mathbb{R}^N)} \int_{\mathbb{R}^N} |Du| + \frac{1}{2\lambda}\int_{\mathbb{R}^N} (u-\chi_C)^2\,dx\,. \tag{24}$$

The following result was proved in [5].

Proposition 2. *We have that* $0 \le u_{\lambda,C} \le 1$, $u_{\lambda,C} = 0$ *in* $\mathbb{R}^N \setminus C$, *and* $u_{\lambda,C}$ *is concave in* $\{u_{\lambda,C} > 0\}$.

The proof of $0 \le u_{\lambda,C} \le 1$ follows from a weak version of the maximum principle [5]. Thanks to the convexity of C, by comparison with the characteristic function of hyperplanes, one can show that $u_{\lambda,C} = 0$ out of C [5]. To prove that $u_{\lambda,C}$ is concave in $\{u_{\lambda,C} > 0\}$ one considers first the case where C is of class $C^{1,1}$ and $\lambda > 0$ is small enough. Then one proves that $u_{\lambda,C}$ is concave by approximating $u_{\lambda,C}$ by the solution u_ϵ of

$$\begin{aligned} &u - \lambda \mathrm{div}\left(\frac{\nabla u}{\sqrt{\epsilon^2 + |\nabla u|^2}}\right) && \text{in } C \\ &\frac{\nabla u}{\sqrt{\epsilon^2 + |\nabla u|^2}}\cdot \nu^C = 0 && \text{in } \partial C, \end{aligned} \tag{25}$$

as $\epsilon \to 0+$, using Korevaar's concavity Theorem [46]. Then one considers the case where C is of class $C^{1,1}$ and we take any $\lambda > 0$. In this case, the concavity of $u_{\lambda,C}$ in $\{u_{\lambda,C} > 0\}$ is derived after proving Theorems 5 and 6 below. The final step proceeds by approximating a general convex body C by convex bodies of class $C^{1,1}$ [4].

Moreover, since $u_{\lambda,C} = 0$ out of C, the upper level set $\{u_{\lambda,C} > s\} \subseteq C$ for any $s \in (0,1]$. Then, as in Proposition 1, one can prove that for any $s \in (0,1]$ the level set $\{u_{\lambda,C} > s\}$ is a solution of

$$(P)_\mu \qquad \min_{E\subseteq C} P(E) - \mu|E|. \tag{26}$$

for the value of $\mu = \lambda^{-1}(1-s)$. When taking $\lambda \in (0,+\infty)$ and $s \in (0,1]$ we are able to cover the whole range of $\mu \in [0,\infty)$ [5]. By Lemma 1 we know that if $\mu < \mu'$ and $C_\mu, C_{\mu'}$ are minimizers of $(P)_\mu$, $(P)_{\mu'}$, respectively, then $C_\mu \subseteq C_{\mu'}$. This implies that the solution of $(P)_\mu$ is unique for any value $\mu \in (0,\infty)$ up to a countable exceptional set. Thus the sets C_μ can be identified with level sets of $u_{\lambda,C}$ for some $\lambda > 0$ and, therefore, we obtain its uniqueness from the concavity of $u_{\lambda,C}$. One can prove [4,5,21]:

Theorem 5. *There is a value $\mu^* > 0$ such that*

$$\begin{cases} \text{if } \mu < \mu^*,\ C_\mu = \emptyset, \\ \text{if } \mu > \mu^*,\ C_\mu \text{ is unique (and convex)}, \\ \text{if } \mu = \mu^*,\ \text{there are two solutions } \emptyset \text{ and } C_{\mu^*}, \end{cases}$$

where C_{μ^*} is the unique Cheeger set of C. Moreover for any $\lambda < \|\chi_C\|_*$ we have $\mu^* := \frac{1-\|u_{\lambda,C}\|_\infty}{\lambda}$ and $C_\mu := \{u_{\lambda,C} > 1 - \mu\lambda\}$ for any $\mu > \mu^*$, where

$$\|\chi_C\|_* := \max\left\{ \int_{\mathbb{R}^N} u\,\chi_C\,dx :\ u \in \mathrm{BV}(\mathbb{R}^N),\ \int_{\mathbb{R}^N} |Du| \le 1 \right\}.$$

The set C_{μ^*} coincides with the level set $\{u_{\lambda,C} = \|u_{\lambda,C}\|_\infty\}$ and is of class $C^{1,1}$.

We call a Cheeger set in a nonempty open bounded subset Ω of $\mathbb{R}^N$ any set $G \subseteq \Omega$ which minimizes

$$\mathcal{C}_\Omega := \min_{F \subseteq \Omega} \frac{P(F)}{|F|}. \tag{27}$$

The Theorem contains the assertion that there is a unique Cheeger set in any nonempty convex body of $\mathbb{R}^N$ and $\mu^* = \mathcal{C}_\Omega$. This result was proved in [21] for uniformly convex bodies of class C^2, and in [4] in the general case. Notice that the solution of (24) gives a practical algorithm to compute the Cheeger set of C.

Theorem 6. *Let C be a nontrivial convex body in $\mathbb{R}^N$. Let*

$$H_C(x) := \begin{cases} -\inf\{\mu : x \in C_\mu\} & \text{if} \quad x \in C \\ 0 & \text{if} \quad x \in \mathbb{R}^N \setminus C. \end{cases}$$

Then $u_{\lambda,C}(x) := (1 + \lambda H_C(x))^+ \chi_C$.

If $N = 2$ and $\mu > \mu^*$, the set C_μ coincides with the union of all balls of radius $1/\mu$ contained in C [6]. Thus its boundary outside ∂C is made by arcs of circle which are tangent to ∂C. In particular, if C is a square, then the Cheeger set corresponds to the arcs of circle with radius $R > 0$ such that $\frac{P(C_{\mu^*})}{|C_{\mu^*}|} = \frac{1}{R}$. We can see that the corners of C are rounded and the discontinuity disappears as soon as $\lambda > 0$ (see the left part of Fig. 1). This is a general fact at points of ∂C where its mean curvature is infinite.

Remark 1. By adapting the proof of Proposition 4 in [5] one can prove the following result. If Ω is a bounded subset of $\mathbb{R}^N$ with Lipschitz continuous boundary, and $u \in \mathrm{BV}(\Omega) \cap L^2(\Omega)$ is the solution of the variational problem

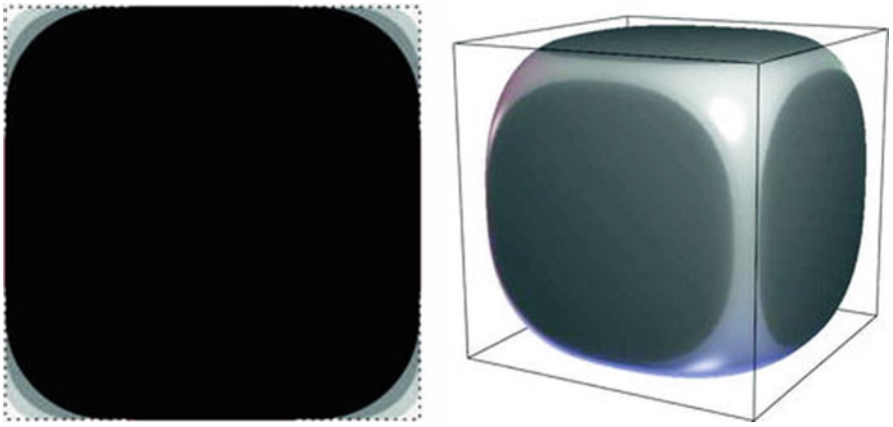


Fig. 1 *Left*: The denoising of a square. *Right*: The Cheeger set of a cube

$$\min_{u\in \mathrm{BV}(\Omega)\cap L^2(\Omega)} \left\{ \int_\Omega |Du| + \frac{1}{2\lambda}\int_\Omega (u-1)^2\,dx + \int_{\partial\Omega} |u|\,d\mathcal{H}^{N-1} \right\}, \tag{28}$$

then $0 \le u \le 1$ and for any $s \in (0,1]$ the upper level set $\{u \ge s\}$ is a solution of

$$\min_{F\subseteq\Omega} P(F) - \lambda^{-1}(1-s)|F|. \tag{29}$$

If $\lambda > 0$ is big enough, indeed greater than $1/\|\chi_\Omega\|_*$, then the level set $\{u = \|u\|_\infty\}$ is the maximal Cheeger set of Ω. In particular, the maximal Cheeger set can be computed by solving (28), and for that we can use the algorithm in [25] described in section "Chambolle's Algorithm." In the right side of Fig. 1 we display the Cheeger set of a cube.

Other explicit solutions corresponding to the union of convex sets can be found in [2, 13]. In particular, Allard [2] describes the solution corresponding to the union of two disks in the plane and also the case of two squares with parallel sides touching by a vertex. Some explicit solutions for functions whose level sets are a finite number of convex sets in $\mathbb{R}^2$ can be found in [14].

5 Numerical Algorithms: Iterative Methods

Notation

Let us fix our main notations. We denote by X the Euclidean space $\mathbb{R}^{N\times N}$. The Euclidean scalar product and the norm in X will be denoted by $\langle\cdot,\cdot\rangle_X$ and $\|\cdot\|_X$,

respectively. Then the image $u \in X$ is the vector $u = (u_{i,j})_{i,j=1}^N$, and the vector field ξ is the map $\xi : \{1, \dots, N\} \times \{1, \dots, N\} \to \mathbb{R}^2$. To define the discrete total variation, we define a discrete gradient operator. If $u \in X$, the discrete gradient is a vector in $Y = X \times X$ given by

$$\nabla u := (\nabla_x u, \nabla_y u),$$

where

$$(\nabla_x u)_{i,j} = \begin{cases} u_{i+1,j} - u_{i,j} & \text{if } i < N \\ 0 & \text{if } i = N, \end{cases} \tag{30}$$

$$(\nabla_y u)_{i,j} = \begin{cases} u_{i,j+1} - u_{i,j} & \text{if } j < N \\ 0 & \text{if } j = N \end{cases} \tag{31}$$

for $i, j = 1, \dots, N$. Notice that the gradient is discretized using forward differences and $\nabla^{+,+}u$ could be a more explicit notation. For simplicity we have preferred to use ∇u. Other choices of the gradient are possible; this one will be convenient for the developments below.

The Euclidean scalar product in Y is defined in the standard way by

$$\langle \xi, \tilde{\xi} \rangle_Y = \sum_{1 \le i,j \le N} (\xi^1_{i,j} \tilde{\xi}^1_{i,j} + \xi^2_{i,j} \tilde{\xi}^2_{i,j})$$

for every $\xi = (\xi^1, \xi^2), \tilde{\xi} = (\tilde{\xi}^1, \tilde{\xi}^2) \in Y$. The norm of $\xi = (\xi^1, \xi^2) \in Y$ is, as usual, $\|\xi\|_Y = \langle \xi, \xi \rangle_Y^{1/2}$. We denote the Euclidean norm of a vector $v \in \mathbb{R}^2$ by $|v|$. Then the discrete total variation is

$$J_d(u) = \|\nabla u\|_Y = \sum_{1 \le i,j \le N} |(\nabla u)_{i,j}|. \tag{32}$$

We have

$$J_d(u) = \sup_{\xi \in Y, |\xi_{i,j}| \le 1 \ \forall (i,j)} \langle \xi, \nabla u \rangle_Y. \tag{33}$$

By analogy with the continuous setting, we introduce a discrete divergence div as the dual operator of ∇, i.e., for every $\xi \in Y$ and $u \in X$ we have

$$\langle -\operatorname{div} \xi, u \rangle_X = \langle \xi, \nabla u \rangle_Y.$$

One can easily check that div is given by

$$(\operatorname{div}\xi)_{i,j} = \begin{cases} \xi^1_{i,j} - \xi^1_{i-1,j} & \text{if } 1 < i < N \\ \xi^1_{i,j} & \text{if } i = 1 \\ -\xi^1_{i-1,j} & \text{if } i = N \end{cases} + \begin{cases} \xi^2_{i,j} - \xi^2_{i,j-1} & \text{if } 1 < j < N \\ \xi^2_{i,j} & \text{if } j = 1 \\ -\xi^2_{i,j-1} & \text{if } j = N \end{cases} \tag{34}$$

for every $\xi = (\xi^1, \xi^2) \in Y$.

We have

$$J_d(u) := \max_{\xi \in \mathcal{V}} \langle u, \operatorname{div}\xi \rangle, \tag{35}$$

where

$$\mathcal{V} = \{\xi \in Y : |\xi_{i,j}|^2 - 1 \le 0, \ \forall\, i, j \in \{1, \dots, N\}\}$$

Chambolle's Algorithm

Let us describe the dual formulation for solving the problem:

$$\min_{u \in X} \quad J_d(u) + \frac{1}{2\lambda}\|u - f\|_X^2 \tag{36}$$

where $f \in X$. Using (35) we have

$$\begin{aligned} \min_{u \in X} J_d(u) + \frac{1}{2\lambda}\|u - f\|_X^2 &= \min_{u \in X} \max_{\xi \in \mathcal{V}} \langle u, \operatorname{div}\xi \rangle + \frac{1}{2\lambda}\|u - f\|_X^2 \\ &= \max_{\xi \in \mathcal{V}} \min_{u \in X} \langle u, \operatorname{div}\xi \rangle + \frac{1}{2\lambda}\|u - f\|_X^2. \end{aligned}$$

Solving explicitly the minimization in u, we have $u = f - \lambda \operatorname{div}\xi$. Then

$$\begin{aligned} \max_{\xi \in \mathcal{V}} \min_{u \in X} \langle u, \operatorname{div}\xi \rangle + \frac{1}{2\lambda}\|u - f\|_X^2 &= \max_{\xi \in \mathcal{V}} \langle f, \operatorname{div}\xi \rangle - \frac{\lambda}{2}\|\operatorname{div}\xi\|_X^2 \\ &= -\frac{\lambda}{2} \min_{\xi \in \mathcal{V}} \left(\left\| \operatorname{div}\xi - \frac{f}{\lambda} \right\|_X^2 - \left\| \frac{f}{\lambda} \right\|_X^2 \right). \end{aligned}$$

Thus if ξ^* is the solution of

$$\min_{\xi\in\mathcal{V}} \left\| \operatorname{div}\xi - \frac{f}{\lambda} \right\|_X^2. \tag{37}$$

then $u = f - \lambda \operatorname{div}\xi^*$ is the solution of (36).

Notice that $\operatorname{div}\xi^*$ is the projection of $\frac{f}{\lambda}$ onto the convex set

$$K_d := \{\operatorname{div}\xi : |\xi_{i,j}| \le 1,\ \forall\, i,j \in \{1,\dots,N\}\}.$$

As in [25], the Karush–Kuhn–Tucker Theorem yields the existence of Lagrange multipliers $\alpha_{i,j} \ge 0$ for the constraints $\xi \in \mathcal{V}$, such that we have for each $(i,j) \in \{1,\dots,N\}^2$

$$\nabla[\operatorname{div}\xi - \lambda^{-1} f]_{i,j} - \alpha^*_{i,j}\xi_{i,j} = 0, \tag{38}$$

with either $\alpha^*_{i,j} > 0$ and $|\xi_{i,j}| = 1$, or $\alpha^*_{i,j} = 0$ and $|\xi_{i,j}| \le 1$. In the later case, we have $\nabla[\operatorname{div}\xi - \lambda^{-1} f]_{i,j} = 0$. In any case, we have

$$\alpha^*_{i,j} = |\nabla[\operatorname{div}\xi - \lambda^{-1} f]_{i,j}|. \tag{39}$$

Let $\nu > 0$, $\xi^0 = 0$, $p \ge 0$. We solve (38) using the following gradient descent (or fixed point) algorithm:

$$\xi^{p+1}_{i,j} = \xi^p_{i,j} + \nu\nabla[\operatorname{div}\xi^p - \lambda^{-1} f]_{i,j} - \nu|\nabla[\operatorname{div}\xi^p - \lambda^{-1} f]_{i,j}|\xi^{p+1}_{i,j}; \tag{40}$$

hence

$$\xi^{p+1}_{i,j} = \frac{\xi^p_{i,j} + \nu\nabla[\operatorname{div}\xi^p - \lambda^{-1} f]_{i,j}}{1 + \nu|\nabla[\operatorname{div}\xi^p - \lambda^{-1} f]_{i,j}|}. \tag{41}$$

Observe that $|\xi^p_{i,j}| \le 1$ for all $i,j \in \{1,\dots,N\}$ and every $p \ge 0$.

Theorem 7. *In the discrete framework, assuming that $\nu < \frac{1}{8}$, then $\operatorname{div}\xi^p$ converges to the projection of $\frac{f}{\lambda}$ onto the convex set K_d. If $\operatorname{div}\xi^*$ is that projection, then $u = f - \lambda\operatorname{div}\xi^*$ is the solution of (36).*

In Fig. 2 we display some results obtained using Chambolle's algorithm with different set of parameters, namely, $\lambda = 5, 10$.

Today, the algorithms Nesterov [49], Beck and Teboulle [12], or the primal-dual approaches described in the next section provide more efficient ways to solve this dual problem.

Fig. 2 Denoising results obtained with Chambolle's algorithm. (**a**) *Top left*: the original image. (**b**) *Top right*: the image with a Gaussian noise of standard deviation $\sigma = 10$. (**c**) *Bottom left*: the result obtained with $\lambda = 5$. (**d**) *Bottom right*: the result obtained with $\lambda = 10$

Primal-Dual Approaches

The primal gradient descent formulation is based on the solution of (36). The dual gradient descent algorithm corresponds to (41). The primal-dual formulation is based on the formulation

$$\min_{u \in X} \max_{\xi \in \mathcal{V}} G(u, \xi) := \langle u, \operatorname{div} \xi \rangle + \frac{1}{2\lambda} \| u - f \|_X^2$$

and performs a gradient descent in u and gradient ascent in ξ.

Given the intermediate solution (u^k, ξ^k) at iteration step k we update the dual variable by solving

$$\max_{\xi \in \mathcal{V}} G(u^k, \xi). \tag{42}$$

Since the gradient ascent direction is $\nabla_\xi G(u^k, \xi) = -\nabla u^k$, we update ξ as

$$\xi^{k+1} = P_{\mathcal{V}}\Big(\xi^k - \frac{\tau_k}{\lambda}\nabla u^k\Big), \tag{43}$$

where τ_k denotes the dual stepsize and $P_{\mathcal{V}}$ denotes the projection onto the convex set $\mathcal{V}$. The projection $P_{\mathcal{V}}$ can be computed as in (41) or simply as

$$(P_{\mathcal{V}}\xi)_{i,j} = \frac{\xi_{i,j}}{\max(|\xi_{i,j}|, 1)}.$$

Now we update the primal variable u by a gradient descent step of

$$\min_{u \in X} G(u, \xi^{k+1}). \tag{44}$$

The gradient ascent direction is $\nabla_u G(u, \xi^{k+1})$ and the update is

$$u^{k+1} = u^k - \theta_k \left(\lambda \operatorname{div} \xi^{k+1} + u^k - f\right), \tag{45}$$

where θ_k denotes the primal stepsize.

The primal-dual scheme was introduced in [57]. The convergence is empirically observed for a variety of suitable stepsize pairs (τ, θ) and is given in terms of the product $\tau\theta$. For instance, convergence is reported for increasing values θ_k and $\tau_k\theta_k \leq 0.5$, see [57]. This has been theoretically explained for a variant of the scheme proposed in [29], while a general convergence proof has been given in [15].

The primal gradient descent and the dual projected gradient descent method are special cases of the above algorithm. Indeed if one solves the problem (42) exactly (taking $\tau_k = \infty$ in (43)) the resulting algorithm is

$$u^{k+1} = u^k - \theta_k \left(-\lambda \operatorname{div} \frac{\nabla u^k}{|\nabla u^k|} + u^k - f\right), \tag{46}$$

with the implicit convention that we may take any element in the unit ball of $\mathbb{R}^2$ when $\nabla u^k = 0$.

If we solve (44) exactly and still apply gradient ascent to (42), the resulting algorithm is

$$\xi^{k+1} = P_{\mathcal{V}}\left(\xi^k + \tau_k \nabla\Big(\operatorname{div} \xi^k - \frac{f}{\lambda}\Big)\right), \tag{47}$$

which essentially corresponds to (41).

The primal-dual approach can be extended to the total variation deblurring problem

$$\min_{u\in X} \quad J_d(u) + \frac{1}{2\lambda}\|Bu - f\|_X^2 \tag{48}$$

where $f \in X$ and B is a matrix representing the discretization of the blurring operator H.

The primal-dual scheme is based on the formulation

$$\min_{u\in X}\max_{\xi\in\mathcal{V}} \quad \langle u, \operatorname{div}\xi\rangle + \frac{1}{2\lambda}\|Bu - f\|_X^2, \tag{49}$$

and the numerical scheme can be written as

$$\begin{aligned} \xi^{k+1} &= P_{\mathcal{V}}\left(\xi^k - \tau_k \nabla u^k\right) \\ u^{k+1} &= u^k - \theta^k\left(-\operatorname{div}\xi^{k+1} + \lambda B^t(Bu^{k+1} - f)\right). \end{aligned} \tag{50}$$

Since B is the matrix of a convolution operator, the second equation can be solved explicitly using the FFT. Convergence is empirically observed for a variety of suitable stepsize pairs (τ, θ) and is given in terms of the product $\tau\theta$, see [57] and a proof in [15]. See also [29] for an explicit rule for acceleration, with proofs of convergence. For a detailed study of different primal-dual methods we refer to [37].

6 Numerical Algorithms: Maximum-Flow Methods

It has been noticed probably first in [51] that Maximal-flow/minimum-cut techniques could be used to solve discrete problems of the form (14), that is, to compute finite sets minimizing a discrete variant of the perimeter and an additional external field term. Combined with (a discrete equivalent of) Proposition 1, this leads to efficient techniques for solving (only) the denoising problem (8), including a method, due to D. Hochbaum, to compute an exact solution in polynomial time (up to machine precision). A slightly more general problem is considered in [27], where the authors describe in detail algorithms which solve the problem with an arbitrary precision.

Discrete Perimeters and Discrete Total Variation

We will call a discrete total variation any convex, nonnegative function $J : \mathbb{R}^M \to [0, +\infty]$ satisfying a discrete *co-area* formula:

$$J(u) = \int_{-\infty}^{+\infty} J(\chi^{\{u\geq s\}})\, ds \tag{51}$$

where $\chi^{\{u\geq s\}} \in \{0,1\}^M$ denotes the vector such that $\chi_i^{\{u\geq s\}} = 0$ if $u_i \leq s$ and $\chi_i^{\{u\geq s\}} = 1$ if $u_i \geq s$.

As an example we can consider the (anisotropic) discrete total variation

$$J(u) = \sum_{\substack{1\leq i<N\\ 1\leq j\leq N}} |u_{i+1,j} - u_{i,j}| + \sum_{\substack{1\leq i\leq N\\ 1\leq j<N}} |u_{i,j+1} - u_{i,j}| \tag{52}$$

In this case $u = (u_{i,j})_{i,j=1}^N$ can be written as a vector in $\mathbb{R}^M$ with $M = N^2$. Then, (51) obviously holds since for any $a, b \in \mathbb{R}$, we have $|a-b| = \int_{-\infty}^{+\infty} |\chi_{\{a>s\}} - \chi_{\{b>s\}}|\, ds$.

Observe, on the other hand, that the discretization (36) does not enter this category (unfortunately). In fact, a discrete total variation will be always very anisotropic (or "crystalline").

We assume that J is not identically $+\infty$. Then, we can derive from (51) the following properties [27]:

Proposition 3. *Let J be a discrete total variation. Then:*

1. *J is positively homogeneous: $J(\lambda u) = \lambda J(u)$ for any $u \in \mathbb{R}^M$ and $\lambda \geq 0$.*
2. *J is invariant by addition of a constant: $J(c\mathbf{1} + u) = J(u)$ for any $u \in \mathbb{R}^M$ and $c \in \mathbb{R}$, where $\mathbf{1} = (1,\dots,1) \in \mathbb{R}^M$ is a constant vector. In particular, $J(\mathbf{1}) = 0$.*
3. *J is lower semicontinuous.*
4. *$p \in \partial J(u) \Leftrightarrow (\forall z \in \mathbb{R},\ p \in \partial J(\chi^{\{u\geq z\}})$.*
5. *J is* submodular*: for any $u, u' \in \{0,1\}^M$,*

$$J(u \vee u') + J(u \wedge u') \leq J(u) + J(u'). \tag{53}$$

More generally, this will hold for any $u, u' \in \mathbb{R}^M$.

Conversely, if $J : \{0,1\}^M \to [0,+\infty]$ is a submodular function with $J(0) = J(\mathbf{1}) = 0$, then the co-area formula (51) extends it to $\mathbb{R}^M$ into a convex function, hence a discrete total variation.

If J is a discrete total variation, then the discrete counterpart of Proposition 1 holds:

Proposition 4. *Let J be a discrete total variation. Let $f \in \mathbb{R}^M$ and let $u \in \mathbb{R}^M$ be the (unique) solution of*

$$\min_{u\in\mathbb{R}^M} \lambda J(u) + \frac{1}{2}\|u - f\|^2 \tag{54}$$

Then, for all $s > 0$, the characteristic functions of the super-level sets $E_s = \{u \geq s\}$ and $E'_s = \{u > s\}$ (which are different only if $s \in \{u_i, i = 1,\dots,M\}$) are

respectively the largest and smallest minimizer of

$$\min_{\theta\in\{0,1\}^M} \lambda J(\theta) \; + \; \sum_{i=1}^{M} \theta_i (s - f_i)\,. \tag{55}$$

The proof is quite clear, since the only properties which were used for showing Proposition 1 were (a) the co-area formula of Theorem 1 and (b) the submodularity of the perimeters (10).

As a consequence, Problem (54) can be solved by successive minimizations of (55), which in turn can be done by computing a maximal flow through a graph, as will be explained in the next section. It seems that efficiently solving the successive minimizations has been first proposed in the seminal work of Eisner and Severance [36] in the context of augmenting-path maximum-flow algorithms. It was then developed, analyzed, and improved by Gallo, Grigoriadis, and Tarjan [39] for preflow-based algorithms. Successive improvements were also proposed by Hochbaum [42], specifically for the minimization of (54). We also refer to [26, 33] for variants, and to [45] for detailed discussions about this approach.

Graph Representation of Energies for Binary MRF

It was first observed by Picard and Ratliff [51] that binary Ising-like energies, that is, of the form

$$\sum_{i,j} \alpha_{i,j} |\theta_i - \theta_j| \; - \; \sum_i \beta_i \theta_i \;, \tag{56}$$

$\alpha_{i,j} \geq 0$, $\beta_i \in \mathbb{R}$, $\theta_i \in \{0,1\}$, could be represented on a graph and minimized by standard optimization techniques, and more precisely using maximum-flow algorithms. Kolmogorov and Zabih [44] showed that the submodularity of the energy is a necessary condition, while, up to sums of ternary submodular interactions, it is also a sufficient condition in order to be representable on a graph. (But other energies are representable, and it does not seem to be known whether any submodular J can be represented on a graph, see [27, Appendix B] and the references therein.)

In case $J(u)$ has only pairwise interactions, as in (52), then Problem (55) has exactly the form (56), with $\alpha_{i,j} = \lambda$ if nodes i and j correspond to neighboring pixels, 0 else, and β_i is $s - f_i$.

Let us build a graph as follows: we consider $\mathcal{V} = \{1,\ldots,M\} \cup \{S\} \cup \{T\}$ where the two special nodes S and T are respectively called the "source" and the "sink." We consider then oriented edges (S,i) and (i,T), $i = 1,\ldots,M$, and (i,j), $1 \leq i,j \leq M$, and to each edge we associate a capacity defined as follows:

$$\begin{cases} c(S,i) \;=\; \beta_i^- & i = 1,\ldots,M\,, \\ c(i,T) \;=\; \beta_i^+ & i = 1,\ldots,M\,, \\ c(i,j) \;=\; \alpha_{i,j} & 1 \leq i,j \leq M\,. \end{cases} \tag{57}$$

Here $\beta_i^+ = \max\{0, \beta_i\}$ and $\beta_i^- = \max\{0, -\beta_i\}$, so that $\beta_i = \beta_i^+ - \beta_i^-$. By convention, we consider there is no edge between two nodes if the capacity is zero. Let us denote by $\mathcal{E}$ the set of edges with nonzero capacity and by $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ the resulting oriented graph.

We then define a "cut" in the graph as a partition of $\mathcal{E}$ into two sets $\mathcal{S}$ and $\mathcal{T}$, with $S \in \mathcal{S}$ and $T \in \mathcal{T}$. The cost of a cut is then defined as the total sum of the capacities of the edges that start on the source side of the cut and land on the sink side:

$$C(\mathcal{S}, \mathcal{T}) = \sum_{\substack{(\mu,\nu)\in\mathcal{E} \\ \mu\in\mathcal{S}, \nu\in\mathcal{T}}} c(\mu, \nu)\,.$$

Then, if we let $\theta \in \{0, 1\}^M$ be the characteristic function of $\mathcal{S} \cap \{1, \ldots, M\}$, we have

$$\begin{aligned} C(\mathcal{S}, \mathcal{T}) &= \sum_{i=1}^{M} (1 - \theta_i)\beta_i^- + \theta_i \beta_i^+ + \sum_{i,j=1}^{M} \alpha_{i,j} (\theta_i - \theta_j)^+ \\ &= \sum_{i,j=1}^{M} \alpha_{i,j} (\theta_i - \theta_j)^+ + \sum_{i=1}^{M} \theta_i \beta_i + \sum_{i=1}^{M} \beta_i^- \end{aligned}$$

If $\alpha_{i,j} = \alpha_{j,i}$ (but other situations are also interesting), this is nothing else than energy (56), up to a constant.

Thus, the problem of finding a minimum of (56) [or (55)] can be reformulated as the problem of finding a minimal cut in the graph. Very efficient algorithms are available, based on a duality result of Ford and Fulkerson [1]. It states that the maximum flow on the graph constrained by the capacities of the edges is equal to the minimal cost of a cut. The problem reduces then to find the maximum flow in the graph. This is precisely defined as follows: starting from S, we "push" a quantity $(x_{\mu,\nu})$ along the oriented edges $(\mu, \nu) \in \mathcal{E}$ of the graph, with the constraint that along each edge,

$$0 \leq x_{\mu,\nu} \leq c(\mu, \nu)$$

and that each "interior" node i must satisfy the flow conservation constraint

$$\sum_{\mu} x_{\mu,i} = \sum_{\mu} x_{i,\mu}$$

(while the source S only sends flow to the network, and the sink T only receives).

It is clear that the total flow $f(x) = \sum_i x_{S,i} = \sum_i x_{i,T}$ which can be sent is bounded from above, and not hard to show that a bound is given by a minimal-cost cut $(\mathcal{S}, \mathcal{T})$. The duality theorem of Ford and Fulkerson expresses the fact that this bound is actually reached by the maximal flow $(x_{\mu,\nu})_{(\mu,\nu)\in\mathcal{E}}$ (which maximizes

$f(x)$), and the partition $(\mathcal{S}, \mathcal{T})$ is obtained by cutting along the saturated edges (μ, ν), where $x_{\mu,\nu} = c_{\mu,\nu}$ while $x_{\nu,\mu} = 0$.

We can find starting from S the first saturated edge along the graph, and cut there, or do the same starting from T and scanning the reverse graph: for $\beta_i = s - f_i$, this will usually give the same solution except for a finite number of levels s, which correspond exactly to the levels $\{u_i : i = 1, \ldots, M\}$ of the solution of (54) and are called the "breakpoints."

Several efficient algorithms are available to compute a maximum flow in polynomial time [1]. Although the time complexity of the algorithm in [16], of Boykov and Kolmogorov, is not polynomial, this algorithm seems to outperform others in terms in time computations, as it is particularly designed for the graphs with low connectivity which arise in image processing.

The idea of a "parametric maximum-flow algorithm" [39] is to reuse the same graph (and the "residual graph" which remains after a run of a max-flow algorithm) to solve problems (55) for increasing values $s \in \{s_0, s_1, \ldots, s_n\}$. This is easily shown to solve (54) up to an arbitrary precision (and in polynomial time, see [39]). It seems this idea was already present in a paper of Eisner and Severance [36].

However, it was shown in [42] by D. Hochbaum that in fact the *exact* solution to (54) can be computed, also in polynomial time. Let us now explain the basic idea of this approach; for details we refer to [27, 42].

Let $u = (u_i)_{i=1}^M$ be the (unique) solution of (54). Proposition 4 tells us that as s varies, problem (55) has the same solution $\chi^{\{u \geq s\}}$ as long as s does not cross any of the values $\{u_i : i = 1, \ldots, M\}$, which are precisely the breakpoints.

Assume we have found, for two levels $s_1 < s_2$, solutions $\theta^1 \geq \theta^2$ of (55) and assume also that these solutions differ. It means that there is a breakpoint u_{i_0} in between: there is at least one location i_0 (and possibly other) with $s_1 \leq u_{i_0} \leq s_2$.

Suppose for a while that the value u_{i_0} were the *only* breakpoint between s_1 and s_2 (that is, at no other location i_1, we can have both $s_1 \leq u_{i_1} \leq s_2$ and $u_{i_0} \neq u_{i_1}$).

In this case, for $s \in [s_1, s_2]$, the optimal energy should be

$$\mathcal{F}(s) = \mathcal{F}_1(s) = \left(\lambda J(\theta^1) - \sum_{i=1}^M \theta_i^1 f_i\right) + s \sum_{i=1}^M \theta_i^1$$

if $s \leq u_{i_0}$, and

$$\mathcal{F}(s) = \mathcal{F}_2(s) = \left(\lambda J(\theta^2) - \sum_{i=1}^M \theta_i^2 f_i\right) + s \sum_{i=1}^M \theta_i^2$$

for $s \geq u_{i_0}$. And the value u_{i_0} is the necessary (only) solution of the equation $\mathcal{F}_1(u_{i_0}) = \mathcal{F}_2(u_{i_0})$.

Observe that in any case, as $\theta^1 \geq \theta^2$ and they are different, the slope of the affine function $\mathcal{F}_1(s)$ is strictly above the slope of the affine function $\mathcal{F}_2(s)$. Since also $\mathcal{F}_1(s_1) \leq \mathcal{F}_2(s_1)$ (as θ^1 is optimal for s_1) and $\mathcal{F}_2(s_2) \leq \mathcal{F}_1(s_2)$, there is always a (unique) value $s_3 \in [s_1, s_2]$ for which $\mathcal{F}_1(s_3) = \mathcal{F}_2(s_3)$.

The idea of the algorithm is now clear: we have to compute a new maximal flow (which, in fact, reuses the residual flows from the computations of θ^1 and θ^2) to solve (55) for the level $s = s_3$. We find a solution θ^3, of energy

$$\mathcal{F}_3(s_3) = \left(\lambda J(\theta^3) - \sum_{i=1}^{M} \theta_i^3 f_i\right) + s_3 \sum_{i=1}^{M} \theta_i^3$$

Then, there are two cases:

- Either $\mathcal{F}_3(s_3) = \mathcal{F}_1(s_3) = \mathcal{F}_2(s_3)$: in this case we have found a breakpoint, and there is no other in the interval $[s_1, s_2]$. Hence, the level sets $\{u \geq s\}$ have been found for all values $s \in [s_1, s_2]$: $\chi^{\{u \geq s\}} = \theta^1$ for $s \in [s_1, s_3]$ and θ^2 for $s \in [s_3, s_2]$.
- Or $\mathcal{F}_3(s_3) < \mathcal{F}_1(s_3) = \mathcal{F}_2(s_3)$. Then, in particular, it must be that the solution θ^3 differs from both θ^1 and θ^2 (otherwise the energies would be the same). Hence we can start again to try solving the problem at the levels s_4 and s_5 which solve $\mathcal{F}_1(s_4) = \mathcal{F}_3(s_4)$ and $\mathcal{F}_3(s_5) = \mathcal{F}_2(s_5)$. Now, since there are only a finite number of possible sets θ solving (55) (bounded by M, as the solutions are nonincreasing with s), this situation can occur at most a finite number of times, bounded by M.

In practice, this can be done in a very efficient way, using "residual graphs" to start the new maximal-flow algorithms, and to compute efficiently the new levels where to cut (there is no need, in fact, to compute the values $\lambda J(\theta) + \sum_i \theta_i f_i$ and $\sum_i \theta_i$ for this). See [27, 42] for details.

For experimental results in the case of total variation denoising we refer to [26, 27, 33, 40].

7 Other Problems: Anisotropic Total Variation Models

Global Solutions of Geometric Problems

The theory of anisotropic perimeters developed in [7] permits to extend model (28) to general anisotropic perimeters, including as particular cases the geodesic active contour model with an inflating force [19, 43], and a model for edge linking [22]. This permits to find the global minima of geometric problems that appear in image processing [22, 24, 27, 32].

The Anisotropic Total Variation and Perimeter Let us define the general notion of total variation with respect to an anisotropy. Following [7] we say that a function $\phi : \Omega \times \mathbb{R}^N \to [0, \infty)$ is a *metric integrand* if ϕ is a Borel function satisfying the conditions:

$$\text{for a.e. } x \in \Omega, \text{ the map } \xi \in \mathbb{R}^N \to \phi(x, \xi) \text{ is convex,} \tag{58}$$

$$\phi(x,t\xi) = |t|\phi(x,\xi) \qquad \forall x \in \Omega, \quad \forall \xi \in \mathbb{R}^N, \quad \forall t \in \mathbb{R}, \tag{59}$$

and there exists a constant $\Lambda > 0$ such that

$$0 \le \phi(x,\xi) \le \Lambda\|\xi\| \qquad \forall x \in \Omega, \quad \forall \xi \in \mathbb{R}^N. \tag{60}$$

We could be more precise and use the term symmetric metric integrand, but for simplicity we use the term metric integrand. Recall that the polar function $\phi^0 : \Omega \times \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$ of ϕ is defined by

$$\phi^0(x,\xi^*) = \sup\{\langle \xi^*, \xi\rangle : \xi \in \mathbb{R}^N \, \phi(x,\xi) \le 1\}. \tag{61}$$

The function $\phi^0(x,\cdot)$ is convex and lower semicontinuous.

Let

$$\mathcal{K}_\phi(\Omega) := \{\sigma \in X_\infty(\Omega) : \phi^0(x,\sigma(x)) \le 1 \text{ for a.e. } x \in \Omega, [\sigma \cdot \nu^\Omega] = 0\}.$$

Definition 3. Let $u \in L^1(\Omega)$. We define the ϕ-total variation of u in Ω as

$$\int_\Omega |Du|_\phi := \sup\left\{\int_\Omega u \operatorname{div}\sigma \, dx : \sigma \in \mathcal{K}^\infty_\phi(\Omega)\right\}, \tag{62}$$

We set $\mathrm{BV}_\phi(\Omega) := \{u \in L^1(\Omega) : \int_\Omega |Du|_\phi < \infty\}$ which is a Banach space when endowed with the norm $|u|_{\mathrm{BV}_\phi(\Omega)} := \int_\Omega |u| dx + \int_\Omega |Du|_\phi$.

We say that $E \subseteq \mathbb{R}^N$ has finite ϕ-perimeter in Ω if $\chi_E \in \mathrm{BV}_\phi(\Omega)$. We set

$$P_\phi(E,\Omega) := \int_\Omega |D\chi_E|_\phi.$$

If $\Omega = \mathbb{R}^N$, we denote $P_\phi(E) := P_\phi(E,\mathbb{R}^N)$. By assumption (60), if $E \subseteq \mathbb{R}^N$ has finite perimeter in Ω it has also finite ϕ-perimeter in Ω.

A Variational Problem and Its Connection with Geometric Problems Let $\phi : \Omega \times \mathbb{R}^N \to \mathbb{R}$ be a metric integrand in Ω and $h \in L^\infty(\Omega)$, $h(x) > 0$ a.e., with $\int_\Omega \frac{1}{h(x)} dx < \infty$. Let us consider the problem

$$\min_{u \in \mathrm{BV}_\phi(\Omega)} \int_\Omega |Du|_\phi + \int_{\partial\Omega} \phi(x,\nu^\Omega)|u| \, d\mathcal{H}^{N-1} + \frac{\lambda}{2}\int_\Omega h\,(u-f)^2\,dx, \tag{63}$$

where ν^Ω denotes the outer unit normal to $\partial\Omega$. To shorten the expressions inside the integrals we shall write h, u instead of $h(x), u(x)$, with the only exception of $\phi(x,\nu^\Omega)$. The following result was proved in [22].

Theorem 8. *(i) Let $f \in L^2(\Omega, h\,dx)$, i.e., $\int_\Omega f(x)^2\, h(x)\, dx < \infty$. Then there is a unique solution of the problem (63).*

(ii) If $u \in \mathrm{BV}_\phi(\Omega) \cap L^2(\Omega, h\,dx)$ be the solution of the variational problem (63) with $f = 1$. Then $0 \le u \le 1$ and the level sets $E_s := \{x \in \Omega : u(x) \ge s\}$, $s \in (0, 1]$, are solutions of

$$\min_{F \subseteq \Omega} P_\phi(F) - \mu |F|_h. \tag{64}$$

where $|F|_h = \int_F h(x)\, dx$. As in the Euclidian case, the solution of (64) is unique for any $s \in (0, 1]$ up to a countable exceptional set.

(iii) When λ is big enough, the level set associated with the maximum of u, $\{u = \|u\|_\infty\}$, is the maximal (ϕ, h)-Cheeger set of Ω, i.e., is a minimizer of the problem

$$\inf \left\{ \frac{P_\phi(F)}{|F|_h} : F \subseteq \overline{\Omega} \text{ of finite perimeter, } |F|_h > 0 \right\}. \tag{65}$$

The computation of the maximal (ϕ, h)-Cheeger set [together with the solution of the family of problems (64)] can be computed by adapting Chambolle's algorithm [25] described in section "Chambolle's Algorithm."

Example 1. We illustrate this formalism with two examples: (a) the geodesic active contour model and (b) a model for edge linking.

(a) *The geodesic active contour model.* Let $I : \Omega \to \mathbb{R}^+$ be a given image in $L^\infty(\Omega)$, G be a Gaussian function, and

$$g(x) = \frac{1}{\sqrt{1 + |\nabla (G * I)|^2}}, \tag{66}$$

(where in $G * I$ we have extended I to $\mathbb{R}^N$ by taking the value 0 outside Ω). Observe that $g \in C(\overline{\Omega})$ and $\inf_{x \in \overline{\Omega}} g(x) > 0$. The geodesic active contour model [19, 43] with an inflating force corresponds to the case where $\phi(x, \xi) = g(x)|\xi|$ and $|Du|_\phi = g(x)|Du|$ and $h(x) = 1$, $x \in \Omega$. The purpose of this model is to locate the boundary of an object of the image at the points where the gradient is large. The presence of the inflating term helps to avoid minima collapsing into a point. The model was initially formulated [19, 43] in a level set framework In this case we may write $P_g(F)$ instead of $P_\phi(F)$, and we have $P_g(F) := \int_{\partial^* F} g\, d\mathcal{H}^{N-1}$, where $\partial^* F$ is the reduced boundary of F [9].

In this case the Cheeger sets are a particular instance of geodesic active contour with an inflating force whose constant is $\mu = \mathcal{C}_\Omega^{g,1}$. An interesting feature of this formalism is that it permits to define local Cheeger sets as local (regional) maxima of the function u. They are Cheeger sets in a subdomain of Ω. They can be identified with boundaries of the image and the above formalism

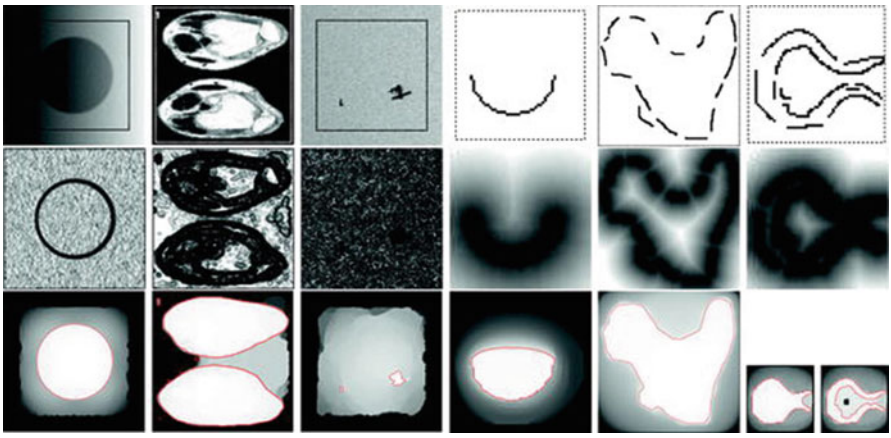


Fig. 3 Geodesic active contours and edge linking experiments. The *first row* shows the images I to be processed. The *first three columns* correspond to segmentation experiments, the *last three* are edge linking experiments. The *second row* shows the weights g used for each experiment (*white* is 1, *black* is 0), in the first two cases $g = (\sqrt{1 + |\nabla(G * I)|^2})^{-1}$, for the third $g = 0.37(\sqrt{0.1 + |\nabla(G * I)|^2})^{-1}$, and for the linking experiments $g = d_S$, the scaled distance function to the given edges. The *third row* shows the disjoint minimum g-Cheeger sets extracted from u (shown in the background); there are 1,7,2,1,1, and 1 sets, respectively. The last linking experiment illustrates the effect of introducing a barrier in the initial domain (*black square*)

permits to compute several active contours at the same time (the same holds true for the edge linking model).

(b) *An edge linking model*. Another interesting application of the above formalism is to edge linking. Given a set $\Gamma \subseteq \Omega$ (which may be curves if $\Omega \subseteq \mathbb{R}^2$ or pieces of surface if $\Omega \subseteq \mathbb{R}^3$), we define $d_\Gamma(x) = \text{dist}(x, \Gamma)$ and the anisotropy $\phi(x, \xi) = d_\Gamma(x)|\xi|$. In that case, we experimentally see that the Cheeger set determined by this anisotropy links the set of curves (or surfaces) Γ. If Γ is a set of edges computed with an edge detector we obtain a set or curves ($N = 2$) or surfaces ($N = 3$) linking them.

Notice that, for a given choice of ϕ, we actually find many *local* ϕ-Cheeger sets, disjoint from the global minimum, that appear as local minima of the Cheeger ratio on the tree of connected components of upper level sets of u. The computation of those sets is partially justified by Proposition 6.11 in [22]. These are the sets which we show on the following experiments.

Let us mention the formulation of active contour models without edges proposed by Chan–Vese in [31] can also be related to the general formulation (64).

In Fig. 3, we display some local ϕ-Cheeger sets of 2D images for the choices of metric ϕ corresponding to geodesic active contour models with an inflating force (the first three columns) and to edge linking problems (the last three columns). The first row displays the original images, and the second row displays the metric

$g = (\sqrt{1+|\nabla(G*I)|^2})^{-1}$ or $g = d_S$. The last row displays the resulting segmentation or set of linked edges, respectively. Let us remark here a limitation of this approach that can be observed in the last subfigure. Even if this linking is produced, the presence of a bottleneck (bottom right subfigure) makes the d_S-Cheeger set to be a set with large volume. This limitation can be circumvented by adding barriers in the domain Ω: we can enforce hard restrictions on the result by removing from the domain some points that we do not want to be enclosed by the output set of curves.

A Convex Formulation of Continuous Multilabel Problems

Let us consider the variational problem

$$\min_{u\in \mathrm{BV}(\Omega),0\le u\le M} \int_\Omega |Du| + \int_\Omega W(x,u(x))\,dx, \tag{67}$$

where $W : \Omega \times \mathbb{R} \to \mathbb{R}^+$ is a potential which is Borel measurable in x and continuous in u, but not necessarily convex. Thus the functional is nonlinear and non-convex. The functional can be relaxed to a convex one by considering the subgraph of u as an unknown.

Our purpose is to write the nonlinearities in (67) in a "convex way" by introducing a new auxiliary variable [52]. This will permit to use standard optimization algorithms. The treatment here will be heuristic.

Without loss of generality, let us assume that $M = 1$. Let $\phi(x,s) = H(u(x)-s)$, where $H = \chi_{[0,+\infty)}$ is the Heaviside function and $s \in \mathbb{R}$. Notice that the set of points where $u(x) > s$ (the subgraph of u) is identified as $\phi(x,s) = 1$. That is, $\phi(x,s)$ is an embedding function for the subgraphs of u. This permits to consider the problem as a binary set problem. The graphs of u is a "cut" in ϕ.

Let

$$\mathcal{A} := \{\phi \in \mathrm{BV}(\Omega\times[0,1]) : \phi(x,s) \in \{0,1\}, \forall (x,s) \in \Omega\times[0,1]\}.$$

Using the definition of anisotropic total variation [7] we may write the energy in (67) in terms of ϕ as

$$\begin{aligned}&\min_{\phi\in\mathcal{A}} \int_\Omega \int_0^1 (|D_x\phi| + W(x,s)|\partial_s\phi(x,s)|)\,dx\,dt + \\ &\int_\Omega (W(x,0)|\phi(x,0)-1| + W(x,1)|\phi(x,1)|)\,dx,\end{aligned} \tag{68}$$

where the boundary conditions $\phi(x,0) = 1$, $\phi(x,1) = 0$ are taken in a variational sense.

Although the energy (68) is convex in ϕ the problem is non-convex since the minimization is carried on $\mathcal{A}$ which is a non-convex set. The proposal in [52] is to relax the variational problem by allowing ϕ to take values in $[0, 1]$. This leads to the following class of admissible functionals:

$$\tilde{\mathcal{A}} := \{\phi \in \mathrm{BV}(\Omega \times [0,1]) : \phi(x,s) \in [0,1],\ \forall (x,s) \in \Omega \times [0,1],\ \phi_s \leq 0\}. \tag{69}$$

The associated variational problem is written as

$$\min_{\phi \in \tilde{\mathcal{A}}} \int_\Omega \int_0^1 (|D_x \phi| + W(x,s)|\partial_s \phi(x,s)|)\ dx\, dt + \int_\Omega (W(x,0)|\phi(x,0) - 1| + W(x,1)|\phi(x,1)|)\, dx. \tag{70}$$

This problem is now convex and can be solved using the dual or primal-dual numerical schemes explained in sections "Chambolle's Algorithm" and "Primal-Dual Approaches." Formally, the level sets of a solution of (70) give solutions of (67). This can be proved using the developments in [7, 22].

In [30] the authors address the problem of convex formulation of multilabel problems with finitely many values including (67) and the case of non-convex neighborhood potentials like the Potts model or the truncated total variation. The general framework permits to consider the relaxation in $\mathrm{BV}(\Omega)$ of functionals of the form

$$F(u) := \int_\Omega f(x, u(x), \nabla u(x))\, dx \tag{71}$$

where $u \in W^{1,1}(\Omega)$ and $f : \Omega \times \mathbb{R} \times \mathbb{R}^N \to [0, \infty[$ be a Borel function such that $f(x, z, \xi)$ is a convex function of ξ for any $(x, z) \in \Omega \times \mathbb{R}^N$ satisfying some coercivity assumption in ξ. Let f^* denote the Legendre–Fenchel conjugate of f with respect to ξ. If

$$K := \{\phi = (\phi^x, \phi^s) : \Omega \times \mathbb{R} \to \mathbb{R}^2 : \phi \text{ is smooth and } f^*(x, s, \phi^x(x,s)) \leq \phi^s(x,s)\}.$$

then the lower semicontinuous relaxation of F is

$$\mathcal{F}(u) = \sup_{\phi \in K} \int_\Omega \int_{\mathbb{R}} \phi \cdot D\chi_{\{(x,s): s < u(x)\}}.$$

Based on this formula one can use a dual or a primal-dual numerical scheme to minimize $\mathcal{F}(u)$ if one knows how to compute the projection onto the convex set K. We refer to [30, 52] for details.

8 Other Problems: Image Restoration

To approach the problem of image restoration from a numerical point of view we shall assume that the image formation model incorporates the sampling process in a regular grid

$$f_{i,j} = (h * u)_{i,j} + n_{i,j}, \qquad (i,j) \in \{1,\dots,N\}^2, \tag{72}$$

where $u : \mathbb{R}^2 \to \mathbb{R}$ denotes the ideal undistorted image, $h : \mathbb{R}^2 \to \mathbb{R}$ is a blurring kernel, f is the observed sampled image which is represented as a function $f : \{1,\dots,N\}^2 \to \mathbb{R}$, and $n_{i,j}$ is, as usual, a white Gaussian noise with zero mean and standard deviation σ.

Let us denote by Ω_N the interval $[0, N[^2$. As we said in the introduction, in order to simplify this problem, we assume that h, u are functions defined in Ω_N and are periodic of period N in each direction. To fix ideas, we assume that $h, u \in L^2(\Omega_N)$, so that $h * u$ is a continuous function in Ω_N and the samples $(h * u)_{i,j}$, $(i,j) \in \{1,\dots,N\}^2$, have sense.

Let us define the discrete functional

$$J_d^\beta(u) = \sum_{1 \le i,j \le N} \sqrt{\beta^2 + |(\nabla u)_{i,j}|^2}, \qquad \beta \ge 0.$$

For any function $w \in L^2(\Omega_N)$, its Fourier coefficients are

$$\hat{w}_{\frac{l}{N},\frac{j}{N}} = \int_{\Omega_N} w(x,y) e^{-2\pi i \frac{(lx+jy)}{N}} \qquad \text{for } (l,j) \in \mathbb{Z}^2.$$

Our plan is to compute a band limited approximation to the solution of the restoration problem for (72). For that we define

$$\mathcal{B} := \left\{ u \in L^2(\Omega_N) : \hat{u} \text{ is supported in } \left\{ -\frac{1}{2} + \frac{1}{N}, \dots, \frac{1}{2} \right\} \right\}.$$

We notice that $\mathcal{B}$ is a finite-dimensional vector space of dimension N^2 which can be identified with X. Both $J(u) = \int_{\Omega_N} |Du|$ and $J_d^0(u)$ are norms on the quotient space $\mathcal{B}/\mathbb{R}$; hence they are equivalent. With a slight abuse of notation we shall indistinctly write $u \in \mathcal{B}$ or $u \in X$.

We shall assume that the convolution kernel $h \in L^2(\Omega_N)$ is such that $\hat{h}$ is supported in $\{-\frac{1}{2} + \frac{1}{N}, \dots, \frac{1}{2}\}$ and $\hat{h}(0,0) = 1$.

In the discrete framework, the ROF model for restoration is

$$\text{Minimize}_{\,u \in X} J_d^\beta(u) \tag{73}$$

$$\text{subject to} \quad \sum_{i,j=1}^{N} |(h * u)_{i,j} - f_{i,j}|^2 \leq \sigma^2 N^2. \tag{74}$$

Notice again that the image acquisition model (1) is only incorporated through a global constraint. In practice, the above problem is solved via the following unconstrained formulation:

$$\min_{u \in X} \max_{\alpha \geq 0} J_d^{\beta}(u) + \frac{\alpha}{2} \left[\frac{1}{N^2} \sum_{i,j=1}^{N} |(h * u)_{i,j} - f_{i,j}|^2 - \sigma^2 \right] \tag{75}$$

where $\alpha \geq 0$ is a Lagrange multiplier. The appropriate value of α can be computed using Uzawa's algorithm [3] so that the constraint (74) is satisfied. Recall that if we interpret α^{-1} as a penalization parameter which controls the importance of the regularization term, and we set this parameter to be small, then homogeneous zones are well denoised while highly textured regions will lose a great part of its structure. On the contrary, if α^{-1} is set to be small, texture will be kept but noise will remain in homogeneous regions. On the other hand, as the authors of [3] observed, if we use the constrained formulation (73)–(74) or, equivalently (75), then the Lagrange multiplier does not produce satisfactory results since we do not keep textures and denoise flat regions simultaneously, and they proposed to incorporate the image acquisition model as a set of local constraints.

Following [3], we propose to replace the constraint (74) by

$$G * (h * u - f)_{i,j} \leq \sigma^2, \qquad \forall (i,j) \in \{1, \ldots, N\}^2, \tag{76}$$

where G is a discrete convolution kernel such that $G_{i,j} > 0$ for all $(i,j) \in \{1, \ldots, N\}^2$. The effective support of G must permit the statistical estimation of the variance of the noise with (76) [3]. Then we shall minimize the functional $J_d^{\beta}(u)$ on X submitted to the family of constraints (76) (plus eventually the constraint $\sum_{i,j=1}^{N} (h * u)_{i,j} = \sum_{i,j=1}^{N} f_{i,j}$). Thus, we propose to solve the optimization problem:

$$\begin{aligned} &\min_{u \in \mathcal{B}} J_d^{\beta}(u) \\ &\text{subject to} \quad G * (h * u - f)_{i,j}^2 \leq \sigma^2 \qquad \forall (i,j). \end{aligned} \tag{77}$$

This problem is well posed, i.e., there exists a solution and is unique if $\beta > 0$ and $\inf_{c \in \mathbb{R}} G * (f - c)^2 > \sigma^2$. In case that $\beta = 0$ and $\inf_{c \in \mathbb{R}} G * (f - c)^2 > \sigma^2$, then $h * u$ is unique. Moreover, it can be solved with a gradient descent approach and Uzawa's method [3].

To guarantee that the assumptions of Uzawa's method hold we shall use a gradient descent strategy. For that, let $v \in X$ and $\gamma > 0$. At each step we have

to solve a problem like

$$\begin{aligned} &\min_{u\in\mathcal{B}} \gamma|u-v|_X^2 + J_d^\beta(u) \\ &\text{subject to } \; G*(h*u-f)_{i,j}^2 \le \sigma^2 \qquad \forall (i,j). \end{aligned} \tag{78}$$

We solve (78) using the unconstrained formulation

$$\min_{u\in X}\max_{\alpha\ge 0} \mathcal{L}^\gamma(u,\{\alpha\};v),$$

where $\alpha = (\alpha_{i,j})_{i,j=1}^N$ and

$$\mathcal{L}^\gamma(u,\{\alpha\};v) = \gamma|u-v|_X^2 + J_d^\beta(u) + \sum_{i,j=1}^N \alpha_{i,j}(G*(h*u-f)_{i,j}^2 - \sigma^2).$$

Algorithm: TV-Based Restoration Algorithm with Local Constraints

1. Set $u_0 = 0$ or, better, $u_0 = f$. Set $n = 0$.
2. Use Uzawa's algorithm to solve the problem

$$\min_{u\in X}\max_{\alpha\ge 0} \mathcal{L}^\gamma(u,\{\alpha\};u^n), \tag{79}$$

that is:

(a) Choose any set of values $\alpha_{i,j}^0 \ge 0$, $(i,j)\in\{1,\dots,N\}^2$, and $u_0^n = u^n$. Iterate from $p=0$ until convergence of α^p the following steps:

(b) With the values of α^p solve DP(γ, u^n):

$$\min_u \mathcal{L}^\gamma(u,\{\alpha^p\};u^n)$$

starting with the initial condition u_p^n. Let u_{p+1}^n be the solution obtained.

(c) Update α in the following way:

$$\alpha_{i,j}^{p+1} = \max(\alpha_{i,j}^p + \rho(G*(h*u_p^n - f)_{i,j}^2 - \sigma^2), 0) \quad \forall (i,j).$$

Let u^{n+1} be the solution of (79). Stop when convergence of u^n.

We notice that, since $\gamma > 0$, Uzawa's algorithm converges if $f \in h*\mathcal{B}$. Moreover, if u^0 satisfies the constraints, then u^n tends to a solution u of (77) as $n\to\infty$ [3].

Finally, to solve problem (79) in Step 2(b) of the Algorithm we use either the extension of Chambolle's algorithm [25] to the restoration case if we use $\beta = 0$, or the quasi-Newton method as in [38] adapted to solve (79) when $\beta > 0$. For more details, we refer to [3, 38] and references therein.

Some Restoration Experiments

To simulate our data we use the modulation transfer function corresponding to SPOT 5 HRG satellite with Hipermode sampling (see [53] for more details):

$$\hat{h}(\eta_1, \eta_2) = e^{-4\pi\beta_1|\eta_1|} e^{-4\pi\alpha\sqrt{\eta_1^2+\eta_2^2}} \mathrm{sinc}(2\eta_1)\mathrm{sinc}(2\eta_2)\mathrm{sinc}(\eta_1), \tag{80}$$

where $\eta_1, \eta_2 \in [-1/2, 1/2]$, $\mathrm{sinc}(\eta_1) = \sin(\pi\eta_1)/(\pi\eta_2)$, $\alpha = 0.58$, and $\beta_1 = 0.14$. Then we filter the reference image given in Fig. 4a with the filter (80) and we add some Gaussian white noise of zero mean and standard deviation σ (in our case $\sigma = 1$, which is a realistic assumption for the case of satellite images [53]) to obtain the image displayed in Fig. 4b.

Figure 5a displays the restoration of the image in Fig. 4b obtained using the Algorithm of last section with $\beta = 0$. We have used a Gaussian function G of radius 6. The mean value of the constraint is $\mathrm{mean}((G * (Ku - f))^2) = 1.0933$ and RMSE $= 7.9862$. Figure 5b displays the function $\alpha_{i,j}$ obtained.

Figure 6 displays some details of the results that are obtained using a single global constraint (74) and show its main drawbacks. Figure 6a corresponds to the result obtained with the Lagrange multiplier $\alpha = 10$ (thus, the constraint (74) is satisfied). The result is not satisfactory because it is difficult to denoise smooth regions and keep the textures at the same time. Figure 6b shows that most textures are lost when using a small value of α ($\alpha = 2$) and Fig. 6c shows that some noise is present if we use a larger value of α ($\alpha = 1,000$). This result is to be compared with the same detail of Fig. 5a which is displayed in Fig. 6d.

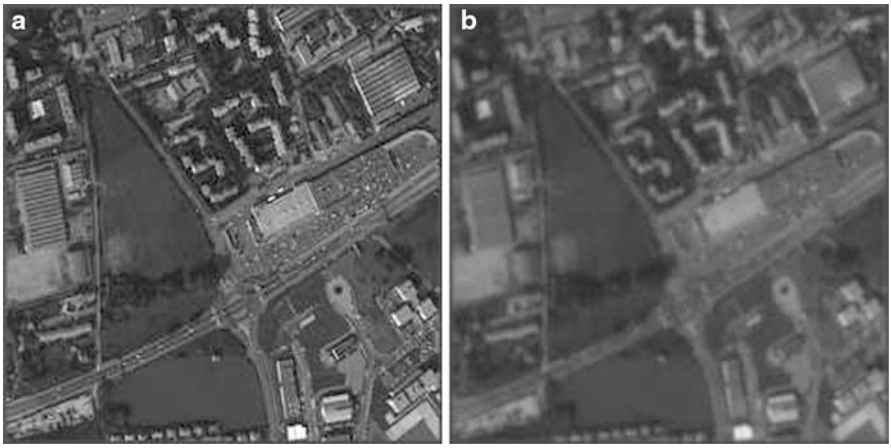


Fig. 4 *Reference image and a filtered and noised image.* (**a**) *Left*: reference image. (**b**) *Right*: the data. This image has been generated applying the MTF given in (80) to the top image and adding a Gaussian white noise of zero mean and standard deviation $\sigma = 1$

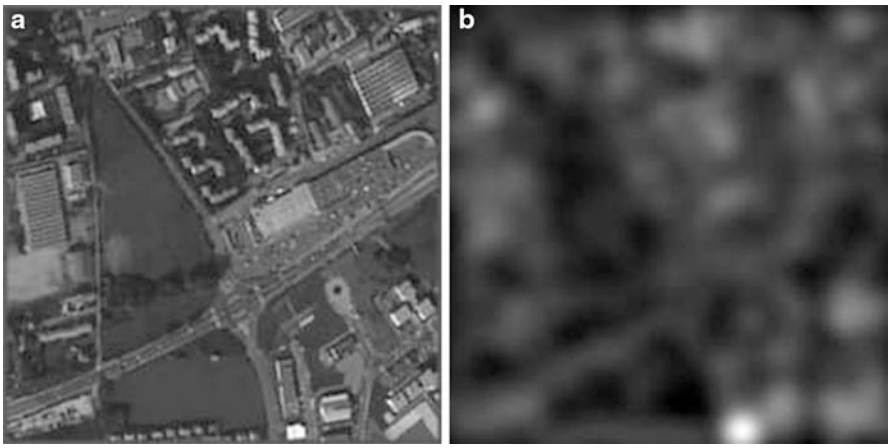


Fig. 5 *Restored image with local Lagrange multipliers.* (**a**) *Left*: the restored image corresponding to the data given in Fig. 4b. The restoration has been obtained using the Algorithm of last section with a Gaussian function G of radius 6. (**b**) *Right*: the function $\alpha_{i,j}$ obtained

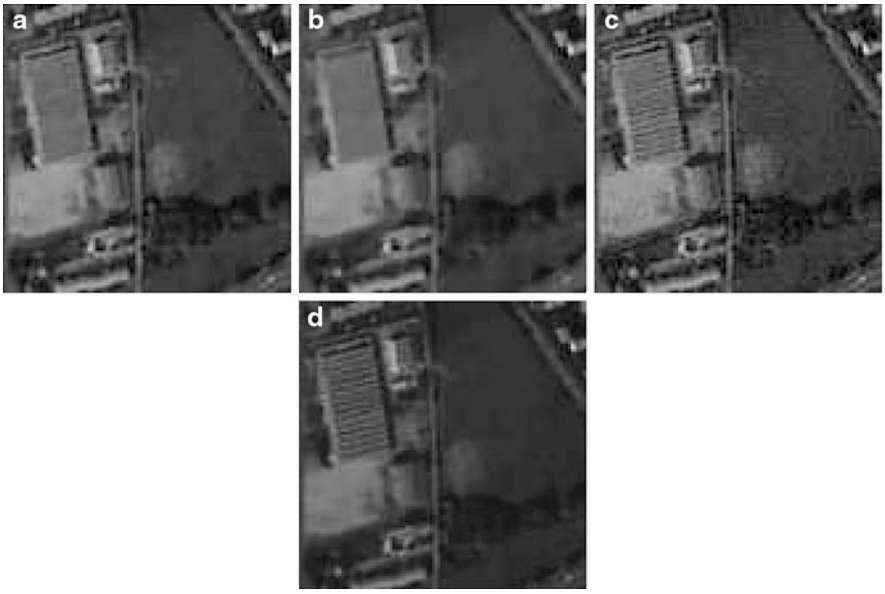


Fig. 6 *A detail of the restored images with global and local constraints. Top*: (**a**), (**b**), and (**c**) display a detail of the results that are obtained using a single global constraint (74) and show its main drawbacks. Figure (**a**) corresponds to the result obtained with the value of α such that the constraint (74) is satisfied, in our case $\alpha = 10$. Figure (**b**) shows that most textures are lost when using a small value of α ($\alpha = 2$) and Figure (**c**) shows that some noise is present if we use a larger value of α ($\alpha = 1{,}000$). *Bottom*: (**d**) displays the same detail of Fig. 5a which has been obtained using restoration with local constraints

The Image Model

For the purpose of image restoration the process of image formation can be modeled in a first approximation by the formula [53]

$$f = Q\{\Pi(h * u) + n\}, \tag{81}$$

where u represents the photonic flux, h is the point spread function of the optical-sensor joint apparatus, Π is a sampling operator, i.e., a Dirac comb supported by the centers of the matrix of digital sensors, n represents a random perturbation due to photonic or electronic noise, and Q is a uniform quantization operator mapping $\mathbb{R}$ to a discrete interval of values, typically $[0, 255]$.

The Modulation Transfer Function for Satellite Images We describe here a simple model for the Modulation Transfer Function of a general satellite. More details can be found in [53] where specific examples of MTF for different acquisition systems are shown. The MTF used in our experiments (80) corresponds to a particular case of the general model described below [53].

Recall that the MTF, which we denote by $\hat{h}$, is the Fourier transform of the point spread function of the system. Let $(\eta_1, \eta_2) \in [-1/2, 1/2]$ denote the coordinates in the frequency domain. There are different parts in the acquisition system that contribute to the global transfer function: the optical system, the sensor, and the blur effects due to motion. Since each subsystem is considered as linear and translation invariant, it is modeled by a convolution operator. The kernel k of the joint system is thus the convolution of the point spread functions of the separated systems.

- **Sensors:** In CCD arrays every sensor has a sensitive region where all the photons that arrive are integrated. This region can be approximated by a unit square $[-c/2, c/2]^2$ where c is the distance between consecutive sensors. Its impulse response is then the convolution of two pulses, one in each spatial direction. The corresponding transfer function also includes the effect of the conductivity (diffusion of information) between neighboring sensors, which is modeled by an exponential decay factor; thus:

$$\hat{h}_S(\eta_1, \eta_2) = \operatorname{sinc}(\eta_1 c)\operatorname{sinc}(\eta_2 c)e^{-2\pi\beta_1 c|\eta_1|}e^{-2\pi\beta_2 c|\eta_2|},$$

  where $\operatorname{sinc}(\eta_1) = \sin(\pi\eta_1)/(\pi\eta_1)$ and $\beta_1, \beta_2 > 0$.
- **Optical system:** The optical system has essentially two effects on the image: it projects the objects from the object plane to the image plane and degrades it. The degradation of the image due to the optical system makes that a light point source loses definition and appears as a blurred (small) region. This effect can be explained by the wave nature of light and its diffraction theory. Discarding other degradation effects due to the imperfect optical systems like lens aberrations [11], the main source of degradation will be the diffraction of the light when passing through a finite aperture: those systems are called diffraction limited systems.

Assuming that the optical system is linear and translation invariant we know that it can be modeled by a convolution operator. Indeed, if the system is linear and translation invariant, it suffices to know the response of the system to a light point source located at the origin, which is modeled by a Dirac delta function δ, since any other light distribution could be approximated (in a weak topology) by superpositions of Dirac functions. The convolution kernel is, thus, the result of the system acting on δ.

If we measure the light intensity and we use a circular aperture the MTF is considered as an isotropic low-pass filter

$$\hat{h}_O(\eta_1, \eta_2) = e^{-2\pi\alpha c\sqrt{\eta_1^2+\eta_2^2}}, \quad \alpha > 0.$$

– **Motion:** each sensor counts the number of photons that arrive to its sensitive region during a certain time of acquisition. During the sampling time the system moves a distance τ and so does the sensor; this produces a motion blur effect in the motion direction (d_1, d_2):

$$\hat{h}_M(\eta_1, \eta_2) = \operatorname{sinc}(\langle(\eta_1, \eta_2), (d_1, d_2)\rangle\tau).$$

Finally, the global MTF is the product of each of these intermediate transfer functions modeling the different aspects of the satellite:

$$\hat{h}(\eta_1, \eta_2) = \hat{h}_S\hat{h}_O\hat{h}_M.$$

Noise We shall describe the typical noise in case of a CCD array. Light is constituted by photons (quanta of light) and those photons are counted by the detector. Typically, the sensor registers light intensity by transforming the number of photons which arrive to it into an electric charge, counting the electrons which the photons take out of the atoms. This is a process of a quantum nature and therefore there are random fluctuations in the number of photons and photoelectrons on the photoactive surface of the detector. To this source of noise we have to add the thermal fluctuations of the circuits that acquire and process the signal from the detector's photoactive surface. This random thermal noise is usually described by a zero-mean white Gaussian process. The photoelectric fluctuations are more complex to describe: for low light levels, photoelectric emission is governed by Bose–Einstein statistics, which can be approximated by a Poisson distribution whose standard deviation is equal to the square root of the mean; for high light levels, the number of photoelectrons emitted (which follows a Poisson distribution) can be approximated by a Gaussian distribution which, being the limit of a Poisson process, inherits the relation between its standard deviation and its mean [11]. In a first approximation this noise is considered as spatially uncorrelated with a uniform power spectrum, thus a white noise. Finally, both sources of noise are assumed to be independent.

Taken together, both sources of noise are approximated by a Gaussian white noise, which is represented in the basic equation (81) by the noise term n. The average signal to noise ratio, called the SNR, can be estimated by the quotient between the signals average and the square root of the variance of the signal.

The detailed description of the noise requires knowledge of the precise system of image acquisition. More details in the case of satellite images can be found in [53] and references therein.

9 Final Remarks: A Different Total Variation-Based Approach to Denoising

Let us briefly comment on the interesting work [47] which interprets the total variation model for image denoising in a Bayesian way leading to a different algorithm based on stochastic optimization which produces better results.

We work again in the discrete setting and consider the image model

$$f_{i,j} = u_{i,j} + n_{i,j} \qquad (i,j) \in \{1,\dots,N\}^2, \tag{82}$$

where $n_{i,j}$ is a white Gaussian noise with zero mean and standard deviation σ.

The solution of (36) can be viewed as a Maximum a Posteriori (MAP) estimate of the original image u. Let $\beta > 0$ and let p_β be the prior probability density function defined by

$$p_\beta(u) \propto e^{-\beta J_d(u)} \qquad u \in X,$$

where we have omitted the normalization constant. The prior distribution models the gradient norms of each pixel as independent and identically distributed random variables following a Laplace distribution. Although the model does not exactly fit the reality since high gradient norms in real images are concentrated along curves and are not independent, it has been found to be convenient and efficient for many tasks in image processing and we follow it here.

Since the probability density of f given u is the density for $n = f - u$, then

$$p(f|u) \propto e^{-\frac{\|f-u\|_X^2}{2\sigma^2}}.$$

Using Bayes rule, the posterior density of u given f is

$$p_\beta(u|f) = \frac{1}{Z} p(f|u) p_\beta(u) = \frac{1}{Z} e^{-\left(\frac{\|f-u\|_X^2}{2\sigma^2} + \beta J_d(u)\right)}, \tag{83}$$

where $Z = \displaystyle\int_{\mathbb{R}^{N^2}} e^{-\left(\frac{\|f-u\|_X^2}{2\sigma^2} + \beta J_d(u)\right)} du$ is the normalization constant making the mass of $p_\beta(u|f)$ to be 1. Then the maximization of the a posteriori density (83) is equivalent to the minimization problem (36) provided that $\beta\sigma^2 = \lambda$.

Fig. 7 (**a**) *Left*: the result obtained by computing $E(u|f)$ and $\beta\sigma^2 = \lambda = 20$, $\sigma = 10$ (image courtesy of Cécile Louchet). (**b**) *Right*: the result obtained using Chambolle's algorithm with $\lambda = 20$

The estimation of u proposed in [47] consists in computing the expected value of u given f:

$$E(u|f) = \frac{1}{Z}\int_{\mathbb{R}^{N^2}} u p_\beta(u|f)\,du = \frac{1}{Z}\int_{\mathbb{R}^{N^2}} u e^{-\left(\frac{\|f-u\|_X^2}{2\sigma^2}+\beta J_d(u)\right)}\,du. \tag{84}$$

This estimate requires to compute an integral in a high dimensional space. In [47], the authors propose to approximate this integral with a Markov Chain Monte-Carlo algorithm (MCMC). In Fig. 7a we display the result of denoising the image in Fig. 2b which has a noise of standard deviation $\sigma = 10$ with the parameter $\beta = \frac{20}{\sigma^2}$. In Fig. 7b we display the denoising of the same image using Chambolle's algorithm with $\lambda = 20$. Notice that in both cases the parameter λ is the same.

10 Conclusion

We have given in this chapter an overview of recent developments on the total variation model in imaging. Its strong influence comes from its ability to recover the image discontinuities and is the basis of numerous applications to denoising, optical flow, stereo imaging and 3D surface reconstruction, segmentation, or interpolation to mention some of them. We have reported the recent theoretical progress on the understanding of its main qualitative features. We have also reviewed the main numerical approaches to solve different models where total variation appears. We have described both the main iterative schemes and the global optimization methods based on the use of max-flow algorithms. Then, we reviewed the use of anisotropic total variation models to solve different geometric problems and its recent use in

finding a convex formulation of some nonconvex total variation problems. We have also studied the total variation formulation of image restoration and displayed some results. We have also reviewed a very recent point of view which interprets the total variation model for image denoising in a Bayesian way, leading to a different algorithm based on stochastic optimization which produces better results.

Acknowledgments We would like to thank Cécile Louchet for providing us the experiments of Sect. 9 and Gabriele Facciolo and Enric Meinhardt for the experiments in section "Global Solutions of Geometric Problems". V. Caselles acknowledges partial support by PNPGC project, reference MTM2006-14836, and also by "ICREA Acadèmia" for excellence in research funded by the Generalitat de Catalunya.

References

1. Ahuja, R.K., Magnanti, T.L., Orlin, J.B.: Network Flows. Theory, Algorithms, and Applications. Prentice Hall, Englewood Cliffs (1993)
2. Allard, W.K.: Total variation regularization for image denoising: I. Geometric theory. SIAM J. Math. Anal. **39**(4), 1150–1190 (2007/2008); Total Variation regularization for image denoising: II. Examples. SIAM J. Imaging Sci. **1**(4), 400–417 (2008)
3. Almansa, A., Ballester, C., Caselles, V., Haro, G.: A TV based restoration model with local constraints. J. Sci. Comput. **34**, 209–236 (2008)
4. Alter, F., Caselles, V.: Uniqueness of the Cheeger set of a convex body. Nonlinear Anal. Theory Methods Appl. **70**, 32–44 (2009)
5. Alter, F., Caselles, V., Chambolle, A.: A characterization of convex calibrable sets in $\mathbb{R}^N$. Math. Ann. **332**, 329–366 (2005)
6. Alter, F., Caselles, V., Chambolle, A.: Evolution of convex sets in the plane by the minimizing total variation flow. Interfaces Free Boundaries **7**, 29–53 (2005)
7. Amar, M., Bellettini, G.: A notion of total variation depending on a metric with discontinuous coefficients. Ann. Inst. H. Poincaré Anal. Non Linéaire **11**(1), 91–133 (1994)
8. Ambrosio, L.: Corso introduttivo alla teoria geometrica della misura ed alle superfici minime. Scuola Normale Superiore, Pisa (1997)
9. Ambrosio, L., Fusco, N., Pallara, D.: Functions of Bounded Variation and Free Discontinuity Problems. Oxford Mathematical Monographs. Oxford University Press, Oxford (2000)
10. Andreu-Vaillo, F., Caselles, V., Mazón, J.M.: Parabolic Quasilinear Equations Minimizing Linear Growth Functionals. Progress in Mathematics, vol. 223. Birkhäuser, Basel (2004)
11. Andrews, H.C., Hunt, B.R.: Digital Signal Processing. Prentice Hall, Englewood Cliffs (1977)
12. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imaging Sci. **2**(1), 183–202 (2009)
13. Bellettini, G., Caselles, V., Novaga, M.: The total variation flow in $\mathbb{R}^N$. J. Differ. Equ. **184**(2), 475–525 (2002)
14. Bellettini, G., Caselles, V., Novaga, M.: Explicit solutions of the eigenvalue problem $-\mathrm{div}\left(\frac{Du}{|Du|}\right) = u$ in $\mathbb{R}^2$. SIAM J. Math. Anal. **36**(4), 1095–1129 (2005)
15. Bonettini, S., Ruggiero, V.: On the convergence of primal-dual hybrid gradient algorithms for total variation image restoration. J. Math. Imaging Vis. **44**(3), 236–253 (2012)
16. Boykov, Y., Kolmogorov, V.: An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. IEEE Trans. Pattern Anal. Mach. Intell. **26**(9), 1124–1137 (2004)
17. Buades, A., Coll, B., Morel, J.M.: A non local algorithm for image denoising. In: Proceedings of the IEEE Conference on CVPR, vol. 2, pp. 60–65 (2005)

18. Caselles, V., Chambolle, A.: Anisotropic curvature-driven flow of convex sets. Nonlinear Anal. **65**(8), 1547–1577 (2006)
19. Caselles, V., Kimmel, R., Sapiro, G.: Geodesic active contours. Int. J. Comput. Vis. **22**(1), 61–79 (1997)
20. Caselles, V., Chambolle, A., Novaga, M.: The discontinuity set of solutions of the TV denoising problem and some extensions. Multiscale Model. Simul. **6**(3), 879–894 (2007)
21. Caselles, V., Chambolle, A., Novaga, M.: Uniqueness of the Cheeger set of a convex body. Pac. J. Math. **232**(1), 77–90 (2007)
22. Caselles, V., Facciolo, G., Meinhardt, E.: Anisotropic Cheeger sets and applications. SIAM J. Imaging Sci. **2**(4), 1211–1254 (2009)
23. Caselles, V., Chambolle, A., Novaga, M.: Regularity for solutions of the total variation denoising problem. Rev. Mat. Iberoam. **27**(1), 233–252 (2011)
24. Chambolle, A.: An algorithm for mean curvature motion. Interfaces Free Boundary **6**(2), 195–218 (2004)
25. Chambolle, A.: An algorithm for total variation minimization and applications. Special issue on mathematics and image analysis. J. Math. Imaging Vis. **20**(1–2), 89–97 (2004)
26. Chambolle, A.: Total variation minimization and a class of binary MRF models. In: Energy Minimization Methods in Computer Vision and Pattern Recognition. Lecture Notes in Computer Science, pp. 136–152. Springer, Berlin (2005)
27. Chambolle, A., Darbon, J.: On total variation minimization and surface evolution using parametric maximum flows. Int. J. Comput. Vis. **84**(3), 288–307 (2009)
28. Chambolle, A., Lions, P.L.: Image recovery via total variation minimization and related problems. Numer. Math. **76**(2), 167–188 (1997)
29. Chambolle, A., Pock, T.: A first-order primal-dual algorithm for convex problems with applications to imaging. J. Math. Imaging Vis. **40**(1), 120–145 (2011)
30. Chambolle, A., Cremers, D., Pock, T.: A convex approach to minimal partitions. SIAM J. Imaging Sci. **5**(4), 1113–1158 (2012)
31. Chan, T.F., Vese, L.A.: Active contours without edges. IEEE Trans. Image Process. **10**(2), 266–277 (2001)
32. Chan, T.F., Esedoglu, S., Nikolova, M.: Algorithms for finding global minimizers of image segmentation and denoising methods. SIAM J. Appl. Math. **66**(5), 1632–1648 (2006)
33. Darbon, J., Sigelle, M.: Image restoration with discrete constrained total variation part I: fast and exact optimization. J. Math. Imaging Vis. **26**(3), 261–276 (2006)
34. De Giorgi, E., Ambrosio, L.: Un nuovo tipo di funzionale del calcolo delle variazioni. Atti Accad. Naz. Lincei Rend. Cl. Sci. Mat Fis. Natur. s.(8) **82**, 199–210 (1988)
35. Demoment, G.: Image reconstruction and restoration: overview of common estimation structures and problems. IEEE Trans. Acoust. Speech Signal Process. **37**(12), 2024–2036 (1989)
36. Eisner, M.J., Severance, D.G.: Mathematical techniques for efficient record segmentation in large shared databases. J. Assoc. Comput. Mach. **23**(4), 619–635 (1976)
37. Esser, E., Zhang, X., Chan, T.: A general framework for a class of first order primal-dual algorithms for convex optimization in imaging science. SIAM J. Imaging Sci. **3**(4), 1015–1046 (2010)
38. Facciolo, G., Almansa, A., Aujol, J.F., Caselles, V.: Irregular to regular sampling, denoising and deconvolution. Multiscale Model. Simul. **7**(4), 1574–1608 (2009)
39. Gallo, G., Grigoriadis, M.D., Tarjan, R.E.: A fast parametric maximum flow algorithm and applications. SIAM J. Comput. **18**(1), 30–55 (1989)
40. Goldfarb, D., Yin, Y.: Parametric maximum flow algorithms for fast total variation minimization. SIAM J. Sci. Comput. **31**(5), 3712–3743 (2009)
41. Gousseau, Y., Morel, J.M.: Are natural images of bounded variation? SIAM J. Math. Anal. **33**(3), 634–648 (2001)
42. Hochbaum, D.S.: An efficient algorithm for image segmentation, Markov random fields and related problems. J. ACM **48**(4), 686–701 (2001)

43. Kichenassamy, S., Kumar, A., Olver, P., Tannenbaum, A., Yezzi, A.: Conformal curvature flows: from phase transitions to active vision. Arch. Ration. Mech. Anal. **134**(3), 275–301 (1996)
44. Kolmogorov, V., Zabih, R.: What energy functions can be minimized via graph cuts? IEEE Trans. Pattern Anal. Mach. Intell. **2**(26), 147–159 (2004)
45. Kolmogorov, V., Boykov, Y., Rother, C.: Applications of parametric maxflow in computer vision. In: Proceedings of the IEEE 11th International Conference on Computer Vision (ICCV 2007), pp. 1–8 (2007)
46. Korevaar, N.: Capillary surface convexity above convex domains. Indiana Univ. Math. J. **32**(1), 73–82 (1983)
47. Louchet, C., Moisan, L.: Total variation denoising using posterior expectation. In: Proceedings of the European Signal Processing Conference (EUSIPCO 2008), Lausanne (2008)
48. Meyer, Y.: Oscillating Patterns in Image Processing and Nonlinear Evolution Equations. The Fifteenth Dean Jacqueline B. Lewis Memorial Lectures. University Lecture Series, vol. 22. American Mathematical Society, Providence (2001)
49. Nesterov, Y.: Smooth minimization of nonsmooth functions. Math. Program. Ser. A **103**(1), 127–152 (2005)
50. Nikolova, M.: Local strong homogeneity of a regularized estimator. SIAM J. Appl. Math. **61**(2), 633–658 (2000)
51. Picard, J.C., Ratliff, H.D.: Minimum cuts and related problems. Networks **5**(4), 357–370 (1975)
52. Pock, T., Schoenemann, T., Cremers, D., Bischof, H.: A convex formulation of continuous multi-label problems. In: European Conference on Computer Vision (ECCV), Marseille (2008)
53. Rougé, B.: Théorie de l'echantillonage et satellites d'observation de la terre. Analyse de Fourier et traitement d'images, Journées X-UPS (1998)
54. Rudin, L., Osher, S.: Total variation based image restoration with free local constraints. In: Proceedings of the IEEE ICIP-94, Austin, vol. 1, pp. 31–35, (1994)
55. Rudin, L., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. Physica D **60**, 259–268 (1992)
56. Scherzer, O., Grasmair, M., Grossauer, H., Haltmeier, M., Lenzen, F.: Variational Methods in Imaging. Applied Mathematical Sciences, vol. 167. Springer, New York (2009)
57. Zhu, M., Chan, T.F.: An efficient primal-dual hybrid gradient algorithm for total variation image restoration. UCLA CAM Report, 08-34 (2008)
58. Ziemer, W.P.: Weakly Differentiable Functions. Sobolev Spaces and Functions of Bounded Variation. Graduate Texts in Mathematics, vol. 120. Springer, New York (1989)

Numerical Methods and Applications in Total Variation Image Restoration

Raymond Chan, Tony F. Chan, and Andy Yip

Contents

1 Introduction 1502
2 Background 1502
3 Mathematical Modeling and Analysis 1503
Variants of Total Variation 1503
Further Applications 1508
4 Numerical Methods and Case Examples 1514
Dual and Primal-Dual Methods 1515
Bregman Iteration 1521
Graph Cut Methods 1524
Quadratic Programming 1527
Second-Order Cone Programming 1528
Majorization-Minimization 1530
Splitting Methods 1532
5 Conclusion 1534
Cross-References 1534
References 1534

Abstract

Since their introduction in a classic paper by Rudin, Osher, and Fatemi (Physica D 60:259–268, 1992), total variation minimizing models have become one of

R. Chan (✉)
Department of Mathematics, The Chinese University of Hong Kong, Shatin, Hong Kong
e-mail: rchan@math.cuhk.edu.hk

T.F. Chan
Office of the President, Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong
e-mail: tonyfchan@ust.hk

A. Yip
Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong
e-mail: mhyipa@hkbu.edu.hk

© Springer Science+Business Media New York 2015

O. Scherzer (ed.), *Handbook of Mathematical Methods in Imaging*,
DOI 10.1007/978-1-4939-0790-8_24

the most popular and successful methodologies for image restoration. New developments continue to expand the capability of the basic method in various aspects. Many faster numerical algorithms and more sophisticated applications have been proposed. This chapter reviews some of these recent developments.

1 Introduction

Images acquired through an imaging system are inevitably degraded in various ways. The types of degradation include noise corruption, blurring, missing values in the pixel domain or transformed domains, intensity saturation, jittering, etc. Such degradations can have adverse effects on high-level image processing tasks such as object detection and recognition. Image restoration aims at recovering the original image from its degraded version(s) to facilitate subsequent processing tasks. Image data differ from many other kinds of data due to the presence of edges, which are important features in human perception. It is therefore essential to preserve and even reconstruct edges in the processing of images. Variational methods for image restoration have been extensively studied in the past couple of decades. A promise of these methods is that the geometric regularity of the resulting images is explicitly controlled by using well-established descriptors in geometry. For example, smoothness of object boundaries can be easily manipulated by controlling their length. There has also been much research in designing variational methods for preserving other important image features such as textures.

Among the various restoration problems, denoising is perhaps the most fundamental one. Indeed, all algorithms for solving ill-posed restoration problems have to have some denoising capabilities either explicitly or implicitly, for otherwise they cannot cope with any error (noise) introduced during image acquisition or numerical computations. Moreover, the noise removal problem boils down to the fundamental problem of modeling natural images which has great impacts on any image processing tasks. Therefore, research on image denoising has been very active.

2 Background

Total variation (TV)-based image restoration models are introduced by Rudin, Osher, and Fatemi (ROF) in their seminal work [51] on edge-preserving image denoising. It is one of the earliest and best-known examples of variational partial differential equation (PDE)-based edge-preserving denoising models. In this model, the geometric regularity of the resulting image is explicitly imposed by reducing the amount of oscillation while allowing for discontinuities (edges). The unconstrained version introduced in [1] reads:

$$\inf_{u \in L^2(\Omega)} \int_\Omega |\nabla u| + \mu \int_\Omega (u - f)^2 \, d\mathbf{x}.$$

Here, Ω is the image domain, $f : \Omega \to \mathcal{R}$ is the observed noisy image, u: $\Omega \to \mathcal{R}$ is the denoised image, and $\mu \geq 0$ is a parameter depending on the noise level. The first term is the *total variation* (TV) which is a measure of the amount of oscillation in the resulting image u. Its minimization would reduce the amount of oscillation which presumably reduces noise. The second term is the L^2 distance between u and f, which encourages the denoised image to inherit most features from the observed data. Thus, the model trades off the closeness to f by gaining the regularity of u. The noise is assumed to be additive and Gaussian with zero mean. If the noise variance level σ^2 is known, then the parameter μ can be treated as the Lagrange multiplier, restraining the resulting image to be consistent with the known noise level, i.e., $\int_\Omega (u - f)^2 d\mathbf{x} = |\Omega|\sigma^2$ [16].

The ROF model is simple and elegant for edge-preserving denoising. Since its introduction, this model has ignited a great deal of research in constructing more sophisticated variants which can give better reconstructed images, designing faster numerical algorithms for solving the optimization problem numerically, and finding new applications in various domains. In a previous book chapter [21] published in 2005, the authors surveyed some recent progresses in the research of total variation-based models. The present chapter aims at highlighting some exciting latest developments in numerical methods and applications of total variation-based methods since the last survey.

3 Mathematical Modeling and Analysis

In this section, the basic definition of total variation and some of its variants are presented. Then, some recent TV-based mathematical models in imaging are reviewed.

Variants of Total Variation

Basic Definition

The use of TV as a regularizer has been shown to be very effective for processing images because of its ability to preserve edges. Being introduced for different reasons, several variants of TV can be found in the literature. Some variants can handle more sophisticated data such as vector-valued imagery and matrix-valued tensors; some are designed to improve restoration quality, and some are modified versions for the ease of numerical implementation. It is worthwhile to review the basic definition and its variants.

In Rudin, Osher, and Fatemi's work [51], the TV of an image $f : \Omega \to \mathcal{R}$ is defined as

$$\int_\Omega |\nabla f| \, d\mathbf{x},$$

where $\Omega \subseteq \mathcal{R}^2$ is a bounded open set. Since the image f may contain discontinuities, the gradient ∇f must be interpreted in a generalized sense. It is well known that elements of the Sobolev space $W^{1,1}(\Omega)$ cannot have discontinuities [2]. Therefore, the TV cannot be defined through the completion of the space C^1 of continuously differentiable functions under the Sobolev norm. The ∇f is thus interpreted as a distributional derivative, and its integral is interpreted as a distributional integral [40]. Under this framework, the minimization of TV naturally leads to a PDE with a distribution as a solution.

Besides defining TV as a distributional integral, other perspectives can offer some unique advantages. A set theoretical way is to define TV as a Radon measure of the domain Ω [50]. This has an advantage of allowing Ω to be a more general set. But a more practical and simple alternative is the "dual formulation." It uses the usual trick in defining weak derivatives – integration by parts – together with the Fenchel transform,

$$\int_\Omega |\nabla f| = \sup \left\{ \int_\Omega f \operatorname{div} \mathbf{g}\, d\mathbf{x} \,\middle|\, \mathbf{g} \in C_c^1\left(\Omega, \mathbb{R}^2\right), |\mathbf{g}(\mathbf{x})| \leq 1 \forall \mathbf{x} \in \Omega \right\} \tag{1}$$

where $f \in L^1(\Omega)$ and div is the divergence operator. Using this definition, one can bypass the discussion of distributions. It also plays an important role in many recent works in dual and primal-dual methods for solving TV minimization problems. The space BV can now be defined as

$$BV(\Omega) := \left\{ f \in L^1(\Omega) \,\middle|\, \int_\Omega |\nabla f| < \infty \right\}.$$

Equipped with the norm $\|f\|_{BV} = \|f\|_{L^1} + \int_\Omega |\nabla f|$, this space is complete and is a proper superset of $W^{1,1}(\Omega)$ [32].

Multichannel TV

Many practical images are acquired in a multichannel way, where each channel emphasizes a specific kind of signal. For example, color images are often acquired through the RGB color components, whereas microscopy images consist of measurements of different fluorescent labels. The signals in the different channels are often correlated (contain redundant information). Therefore, in many practical situations, regularization of multichannel images should not be done independently on each channel.

There are several existing ways to generalize TV to vectorial data. A review of some generalizations can be found in [20]. Many generalizations are very intuitive. But only some of them have a natural dual formulation. Sapiro and Ringach [52] proposed to define

$$\int_\Omega |\nabla f| := \int_\Omega \sqrt{\sum_{i=1}^{M} |\nabla f_i|^2} d\mathbf{x} = \int_\Omega |\nabla f|_F \, d\mathbf{x},$$

where $f = (f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_M(\mathbf{x}))$ is the vectorial data with M channels. Thus, it is the integral of the Frobenius norm $|\cdot|_F$ of the Jacobian ∇f. The dual formulation given in [10] is

$$\sup \left\{ \int_\Omega \langle f, \operatorname{div} \mathbf{g} \rangle d\mathbf{x} \middle| \mathbf{g} \in C_c^1 \left(\Omega, \mathbb{R}^{2\times M}\right), |\mathbf{g}(\mathbf{x})|_F \leq 1 \forall \mathbf{x} \in \Omega \right\},$$

where $\langle f, \operatorname{div} \mathbf{g} \rangle = \sum_{i=1}^{M} f_i \operatorname{div} \mathbf{g}_i$.

Matrix-Valued TV

In applications such as diffusion tensor imaging (DTI), the measurements at each spatial location are represented by a diffusion tensor, which is a 3×3 symmetric positive semi-definite matrix. Recent efforts have been devoted to generalize the TV to matrix-valued images. Some natural generalizations can be obtained by identifying an $M \times N$ matrix with an MN vector, so that a vector-valued total variation can be applied. This was done by Tschumperlé and Deriche [57], who generalized the vectorial TV of [7]. The main challenge is to preserve the positive definiteness of the denoised solution. This will be elaborated in section "Diffusion Tensors Images."

Another interesting approach proposed by Setzer et al. [54] is the so-called operator-based regularization. Given a matrix-valued function $f = (f_{ij}(\mathbf{x}))$, define a matrix function $A := (a_{ij})$ where $a_{ij} = |\nabla f_{ij}|^2$. Let $\Phi(A)$ be the matrix obtained by replacing each eigenvalue λ of A with $\sqrt{\lambda}$. Then the total variation is defined to be $\int_\Omega |\Phi(A)|_F d\mathbf{x}$, where $|\cdot|_F$ is the Frobenius norm. While this formulation seems complicated, its first variation turns out to have a nice simple formula. However, when combined with the ROF model, the preservation of positive definiteness is an issue.

Discrete TV

The ROF model is cast as an infinite-dimensional optimization problem over the BV space. To solve the problem numerically, one must discretize the problem at some stage. The approach proposed by Rudin et al. in [51] is to "optimize then discretize." The gradient flow equation is discretized with a standard finite difference scheme. This method works very well in the sense that the numerical solution converges to a steady state which is qualitatively consistent with the expected result of the (continuous) ROF model. However, to the best of the authors' knowledge, a theoretical proof of convergence of the numerical solution to the exact solution of the gradient flow equation as the grid size tends to zero is not yet available. A standard convergence result of finite difference schemes for nonlinear PDE is based on the compactness of TV-bounded sets in L^1 [46]. However, proving TV boundedness in two or more dimensions is often difficult.

An alternative approach is to "discretize then optimize." In this case, only one has to solve a finite-dimensional optimization problem, whose numerical solution can in many cases be shown to converge. But the convergence of the exact solution

of the finite-dimensional problems to the exact solution of the original infinite-dimensional problem is often hard to obtain too. So, both approaches suffer from the theoretical convergence problem. But the latter method has a precise discrete objective to optimize.

To discretize the ROF objective, the fitting term is often straightforward. But the discretization of the TV term has a strong effect on the numerical schemes. The most commonly used versions of discrete TV are

$$\|f\|_{TV} = \sum_{i=1}^{m-1}\sum_{j=1}^{n-1}\sqrt{\left(f_{i+1,j}-f_{i,j}\right)^2+\left(f_{i,j+1}-f_{i,j}\right)^2}\Delta x, \tag{2}$$

$$\|f\|_{TV} = \sum_{i=1}^{m-1}\sum_{j=1}^{n-1}\left(\left|f_{i+1,j}-f_{i,j}\right|+\left|f_{i,j+1}-f_{i,j}\right|\right)\Delta x, \tag{3}$$

where $f = (f_{i,j})$ is the discrete image and Δx is the grid size. They are sometimes referred as the *isotropic* and *anisotropic* versions, respectively, for they are a formal discretization of the isotropic TV $\int_\Omega \sqrt{f_x^2+f_y^2}d\mathbf{x}$ and the anisotropic TV $\int_\Omega(|f_x|+|f_y|)d\ \mathbf{x}$, respectively. The anisotropic TV is not rotational invariant; an image and its rotation can have a different TV value. Therefore, the discrete TV (3) deviates from the original isotropic TV. But being a piecewise linear function, some numerical techniques for quadratic and linear problems can be applied. Indeed, by introducing some auxiliary variables, the corresponding discrete ROF objective can be converted into a canonical quadratic programming problem [30].

Besides using finite difference approximations, a recent popular way is to represent TV on graphs [27]. To make the problem fully discrete, the range of the image is quantized to a finite set of K integers only, usually 0–255. The image is "leveled," so that $f_{i,j}^k = 1$ if the intensity of the (i,j)th pixel is at most k, and $f_{i,j}^k = 0$ otherwise. Then, the TV is given by

$$\|f\|_{TV} = \sum_{k=0}^{K-1}\sum_{i,j}\sum_{s,t} w_{i,j,s,t}\left|f_{i,j}^k - f_{s,t}^k\right|, \tag{4}$$

where $w_{i,j,s,t}$ is a nonnegative weight. A simple choice is the four-connectivity model, where $w_{i,j,s,t} = 1$ if $|i-s|+|j-t| \le 1$ and $w_{i,j,s,t} = 0$ otherwise. In this case, it becomes the anisotropic TV (3). Different choices of the weights penalize edges in different orientations.

A related concept introduced by Shen and Kang is the *quantum total variation* [55]. They studied the ROF model when the range of an image is a finite discrete set (preassigned or determined on the fly), but the image domain is a continuous one. The model is suitable for problems such as bar code scanning, image quantization, and image segmentation. An elegant analysis of the model and some stochastic gradient descent algorithms were presented there.

Nonlocal TV

First proposed by Buades et al. [11], the nonlocal means algorithm renounces the use of local smoothness to denoise an image. Patches which are spatially far away but photometrically similar are also utilized in the estimation process – a paradigm which has been used in texture synthesis [28]. The denoising results are surprisingly good. Since then, the use of nonlocal information becomes increasingly popular. In particular, Bresson and Chan [10] and Gilboa and Osher [31] considered the nonlocal TV. The nonlocal gradient $\nabla_{NL} f$ for a pair of points $\mathbf{x} \in \Omega$ and $\mathbf{y} \in \Omega$ is defined by

$$\nabla_{NL} f\,(\mathbf{x}, \mathbf{y}) = \sqrt{w\,(\mathbf{x}, \mathbf{y})}\,(f\,(\mathbf{x}) - f\,(\mathbf{y}))\,,$$

where $w(\mathbf{x}, \mathbf{y})$ is a nonnegative weight function which is presumably a similarity measure between a patch around $\mathbf{x}$ and a patch around $\mathbf{y}$. As an illustration, a simple choice of the weight function is

$$w\,(\mathbf{x}, \mathbf{y}) = \alpha_1 e^{-|\mathbf{x}-\mathbf{y}|^2/\sigma_1^2} + \alpha_2 e^{-|F(\mathbf{x})-F(\mathbf{y})|^2/\sigma_2^2},$$

where α_i and σ_i are positive constants, and $F(\mathbf{x})$ is a feature vector derived from a patch around $\mathbf{x}$. The constants α_i may sometimes be defined to depend on $\mathbf{x}$, so that the total weight over all $\mathbf{y} \in \Omega$ is normalized to 1. In this case, the weight function is nonsymmetric with respective to its arguments. The first term in w is a measure of geometric similarity, so that nearby pixels have a higher weight. The second term is a measure of photometric similarity. The feature vector F can be the color histogram or any texture descriptor over a window around $\mathbf{x}$. The norm of the nonlocal gradient at $\mathbf{x}$ is defined by

$$|\nabla_{NL} f\,(\mathbf{x})| = \sqrt{\int_\Omega [\nabla_{NL} f\,(\mathbf{x}, \mathbf{y})]^2\, d\mathbf{y}},$$

which adds up all the squared intensity variation relative to $f(\mathbf{x})$, weighted by the similarity between the corresponding pair of patches. The nonlocal TV is then naturally defined by summing up all the norms of the nonlocal gradients over the image domain:

$$\int_\Omega |\nabla_{NL} f|\, d\mathbf{x}.$$

Therefore, the nonlocal TV is small if, for each pair of similar patches, the intensity difference between their centers is small. An advantage of using the nonlocal TV to regularize images is its tendency to preserve highly repetitive patterns better. In practice, the weight function is often truncated to reduce the computation costs spent in handling the many less similar patches.

Further Applications

Inpainting in Transformed Domains

After the release of the image compression standard JPEG2000, images can be formatted and stored in terms of wavelet coefficients. For instance, in Acrobat 6.0 or later, users can opt to use JPEG2000 to compress embedded images in a PDF file. During the process of storing or transmission, some wavelet coefficients may be lost or corrupted. This prompts the need of restoring missing information in wavelet domains. The setup of the problem is as follows. Denote the standard orthogonal wavelet expansion of the images f and u by

$$f(\alpha) = \sum_{j,k} \alpha_{j,k} \psi_{j,k}(x), \quad j \in \mathbb{Z}, k \in \mathbb{Z}^2,$$

and

$$u(\beta) = \sum_{j,k} \beta_{j,k} \psi_{j,k}(x), \quad j \in \mathbb{Z}, k \in \mathbb{Z}^2,$$

where $\{\Psi_{j,k}\}$ is the wavelet basis, and $\{\alpha_{j,k}\}$, $\{\beta_{j,k}\}$ are the wavelet coefficients of f and u given by

$$\alpha_{j,k} = \langle f, \psi_{j,k} \rangle \quad \text{and} \quad \beta_{j,k} = \langle u, \psi_{j,k} \rangle,$$

respectively, for $j \in \mathbb{Z}$, $k \in \mathbb{Z}^2$. For convenience, $u(\beta)$ is denoted by u when there is no ambiguity. Assume that the wavelet coefficients in the index set I are known, i.e., the available wavelet coefficients are given by

$$\xi_{j,k} = \begin{cases} \alpha_{j,k}, & (j,k) \in I, \\ 0, & (j,k) \in \Omega \backslash I. \end{cases}$$

The aim of wavelet domain inpainting is to reconstruct the wavelet coefficients of u from the given coefficients ξ. It is well known that the inpainting problem is ill posed, i.e., it admits more than one solution. There are many different ways to fill in the missing coefficients, and therefore many different reconstructions in the pixel domain are possible. Regularization methods can be used to incorporate prior information about the reconstruction. In [23], Chan, Shen, and Zhou used TV to solve the wavelet inpainting problem, so that the missing coefficients are filled while preserving sharp edges in the pixel domain faithfully. More precisely, they considered the minimization of the following objective

$$F(\beta) = \frac{1}{2} \sum_{j,k} \chi_{j,k} \left(\xi_{j,k} - \beta_{j,k}\right)^2 + \lambda \|u\|_{TV}, \tag{5}$$

with $\chi_{j,k} = 1$ if $(j, k) \in I$ and $\chi_{j,k} = 0$ if $(j,k) \in \Omega \backslash I$, and λ is the regularization parameter. The first term in F is the data-fitting term, and the second is the TV regularization term. The method Chan, Shen, and Zhou used to optimize the objective is the standard gradient descent. The method is very robust, but it often slows down significantly before it converges.

In [18], Chan, Wen, and Yip proposed an efficient *optimization transfer algorithm to* minimize the objective (5). An auxiliary variable ζ is introduced to yield a new objective function:

$$G(\zeta, \beta) = \frac{1+\tau}{2\tau}\left(\|\chi(\zeta - \xi)\|_2^2 + \tau\|\zeta - \beta\|_2^2\right) + \lambda\|u(\beta)\|_{TV},$$

where χ denotes a diagonal matrix with diagonal entries $\chi_{j,k}$, and τ is an arbitrary positive parameter. The function G is a quadratic majorizing function [43] of F. The method also has a flavor of the splitting methods introduced in section "Splitting Methods." But a major difference is that the method here solves the original problem (5) without any alteration. It can be easily shown that

$$F(\beta) = \min_{\zeta} G(\zeta, \beta)$$

for any positive regularization parameter τ. Thus, the minimization of G w.r.t. (ζ, β) is equivalent to the minimization of F w.r.t. β for any $\tau > 0$. Unlike the gradient descent method of [23], the optimization transfer algorithm avoids the use of derivatives of the TV. It also does not require smoothing out the TV to make it differentiable. The experimental results in [18] showed that the algorithm is very efficient and outperforms the gradient descent method.

Superresolution

Image superresolution refers to the process of increasing spatial resolution by fusing information from a sequence of low-resolution images of the same scene. The images are assumed to contain subpixel information (due to subpixel displacements or blurring), so that the superresolution is possible.

In [24], Chan et al. proposed a unified TV model for superresolution imaging problems. They focused on the problem of reconstructing a high-resolution image from several decimated, blurred, and noisy low-resolution versions of the high-resolution image. They derived a low-resolution image formation model which allows multiple-shifted and blurred low-resolution image frames, so that it subsumes several well-known models. The model also allows an arbitrary pattern of missing pixels (in particular an arbitrary pattern of missing frames). The superresolution image reconstruction problem is formulated as an optimization problem which combines the image formation model and the TV inpainting model. In this method, TV minimization is used to suppress noise amplification, repair corrupted pixels in regions without missing pixels, and reconstruct intensity levels in regions with missing pixels.

Image Formation Model The observation model, Chan et al. considered, consists of various degradation processes. Assume that a number of $m \times n$ low-resolution frames are captured by an array of charge-coupled device (CCD) sensors. The goal is to reconstruct an $Lm \times Ln$ high-resolution image. Thus, the resolution is increased by a factor of L in each dimension. Let u be the ideal $Lm \times Ln$ high-resolution clean image.

1. *Formation of low-resolution frames*. A low-resolution frame is given by

$$D_{p,q} C u,$$

   where C is an averaging filter with window size L-by-L, and $D_{p,q}$ is the downsampling matrix which, starting at the $(p,\ q)$th pixel, samples every other L pixels in both dimensions to form an $m \times n$ image.
2. *Blurring of frames*. This is modeled by

$$H_{p,q} D_{p,q} C u,$$

   where $H_{p,q}$ is the blurring matrix for the $(p,\ q)$th frame.
3. *Concatenation of frames*. The full set of L^2 frames are interlaced to form an $mL \times nL$ image:

$$Au,$$

   where

$$A = \sum_{p,q} D_{p,q}^T H_{p,q} D_{p,q} C.$$

4. *Additive Noise*.

$$Au + \eta,$$

   where each pixel in η is a Gaussian white noise.
5. *Missing pixels and missing frames*.

$$f = \Lambda_{\mathcal{D}} (Au + \eta),$$

   where $\mathcal{D}$ denotes the set of missing pixels, and $\Lambda_{\mathcal{D}}$ is the downsampling matrix from the image domain to $\mathcal{D}$.
6. *Multiple observations*. Finally, multiple observations of the same scene, but with different noise and blurring, are allowed. This leads to the model

$$f_r = \Lambda_{\mathcal{D}_r} (A_r u + \eta_r) \quad r = 1, \ldots, R, \tag{6}$$

where

$$A_r = \sum_{p,q} D_{p,q}^T H_{p,q,r} D_{p,q} C.$$

TV Superresolution Imaging Model To invert the degradation processes in (6), a Tikhonov-type regularization model has been used. It requires minimization of the following energy:

$$F(u) = \frac{1}{2} \sum_{r=1}^{R} \| \Lambda_{\mathcal{D}_r} A_r u - f_r \|^2 + \lambda \| u \|_{TV} .$$

This model simultaneously performs denoising, deblurring, inpainting, and superresolution reconstruction. Experimental results show that reasonably good reconstruction can be obtained even if five-sixth of the pixels are missing and the frames are blurred.

Image Segmentation

TV minimization problems also arise from image segmentation. When one seeks for a partition of the image into homogeneous segments, it is often helpful to regularize the shape of the segments. This can increase the robustness of the algorithm against noise and avoid spurious segments. It may also allow the selection of features of different scales. In the classical Mumford-Shah model [47], the regularization is done by minimizing the total length of the boundary of the segments. In this case, if one represents a segment by its characteristic function, then the length of its boundary is exactly the TV of the characteristic function. Therefore, the minimization of length becomes the minimization of TV of characteristic functions.

Given an observed image f on an image domain Ω, the piecewise constant Mumford-Shah model seeks a set of curves C and a set of constant $\mathbf{c} = (c_1, c_2, \ldots, c_L)$ which minimize the energy functional given by

$$F^{MS}(C, \mathbf{c}) = \sum_{l=1}^{L} \int_{\Omega_l} [f(\mathbf{x}) - c_l]^2 \, d\mathbf{x} + \beta \cdot \text{Length}(C) .$$

The curves in C partition the image into L mutually exclusive segments Ω_l for $l = 1, 2, \ldots, L$. The idea is to partition the image, so that the intensity of f in each segment Ω_l is well approximated by a constant c_l. The goodness of fit is measured by the L^2 difference between f and c_l. On the other hand, a minimum description length principle is employed which requires the curves C to be as short as possible. This increases the robustness to noise and avoids spurious segments. The parameter $\beta > 0$ controls the trade-off between the goodness of fit and the length of the curves C. The Mumford-Shah objective is nontrivial to optimize especially when the curves need to be split and merged. Chan and Vese [24] proposed a level set-based method which can handle topological changes effectively. In the two-phase

version of this method, the curves are represented by the zero level set of a Lipschitz level set function Φ defined on the image domain. The objective function then becomes

$$F^{\mathrm{CV}}(\phi, c_1, c_2) = \int_{\Omega} H(\phi(\mathbf{x}))\,[f(\mathbf{x}) - c_1]^2\,d\mathbf{x} + \int_{\Omega} [1 - H(\phi(\mathbf{x}))]\,[f(\mathbf{x}) - c_2]^2\,d\mathbf{x} + \beta \int_{\Omega} |\nabla H(\phi)|.$$

The function H is the Heaviside function defined by $H(x) = 1$ if $x \geq 0$, $H(x) = 0$ otherwise. In practice, we replace H by a smooth approximation H_ε, e.g.,

$$H_\varepsilon(x) = \frac{1}{2}\left[1 + \frac{2}{\pi}\arctan\left(\frac{x}{\varepsilon}\right)\right].$$

Although this method makes splitting and merging of curves a simple matter, the energy functional is non-convex which possesses many local minima. These local minima may correspond to undesirable segmentations; see [45].

Interestingly, for fixed c_1 and c_2, the above non-convex objective can be reformulated as a convex problem, so that a global minimum can be easily computed; see [22,56]. The globalized objective is given by

$$F^{\mathrm{CEN}}(u, c_1, c_2) = \int_{\Omega} \left\{[f(\mathbf{x}) - c_1]^2 - [f(\mathbf{x}) - c_2]^2\right\} u(\mathbf{x})\,d\mathbf{x} + \beta \int_{\Omega} |\nabla u| \tag{7}$$

which is minimized over all u satisfying the bilateral constraints $0 \leq u \leq 1$ and all scalars c_1 and c_2. After a solution u is obtained, a global solution to the original two-phase Mumford-Shah objective can be obtained by thresholding u with μ for almost every $\mu \in [0,1]$, see [22, 56]. Some other proposals for computing global solutions can be found in [45].

To optimize the globalized objective function (7), Chan et al. [22] proposed to use an exact penalty method to convert the bilaterally constrained problem to an unconstrained problem. Then the gradient descent method is applied. This method is very robust and easy to implement. Moreover, the exact penalty method treats the constraints gracefully, as if there is no constraint at all. But of course the gradient descent is not particular fast.

In [42], Krishnan et al. considered the following discrete two-phase Mumford-Shah model:

$$F^{\mathrm{CEN}}(u, c_1, c_2) = \langle s, u\rangle + \beta\,\|u\|_{TV} + \frac{\alpha}{2}\left\|u - \frac{1}{2}\right\|^2,$$

where $\langle\cdot,\ \cdot\rangle$ is the l^2 inner product, $s = (s_{i,j})$, and

$$s_{i,j} = \left(f_{i,j} - c_1\right)^2 - \left(f_{i,j} - c_2\right)^2.$$

The variable u is bounded by the bilateral constraints $0 \leq u \leq 1$. When $\alpha = 0$, this problem is convex but not strictly convex. When $\alpha > 0$, this problem is strictly convex. The additive constant $\frac{1}{2}$ is introduced in the third term so that the minimizer does not bias toward $u = 0$ or $u = 1$. This problem is exactly a TV denoising problem with bound constraints. Krishnan et al. proposed to use the primal-dual active-set method to solve the problem. Superlinear convergence has been established.

Diffusion Tensors Images

Recently, diffusion tensor imaging (DTI), a kind of magnetic resonance (MR) modality, becomes increasingly popular. It enables the study of anatomical structures such as nerve fibers in human brains noninvasively. Moreover, the use of direction-sensitive acquisitions results in its lower signal-to-noise ratio compared to convectional MR. At each voxel in the imaging domain, the anisotropy of diffusion water molecules is interested. Such an anisotropy can be described by a diffusion tensor D, which is a 3×3 positive semi-definite matrix. By standard spectral theory results, D can be factorized into

$$D = V \Lambda V^T,$$

where V is an orthogonal matrix whose columns are the eigenvectors of D, and Λ is a diagonal matrix whose diagonal entries are the corresponding eigenvalues. These eigenvalues provide the diffusion rate along the three orthogonal directions defined by the eigenvectors. The goal is to estimate the matrix D (one at each voxel) from the data. Under the Stejskal-Tanner model, the measurement S_k from the imaging device and the diffusion tensor are related by

$$S_k = S_0 e^{-b g_k^T D g_k}, \tag{8}$$

where S_0 is the baseline measurement, g_k is the prescribed direction in which the measurement is done, and $b > 0$ is a scalar depending the strength of the magnetic field applied and the acquisition time. Since D has six degrees of freedom, six measurements at different orientations are needed to reconstruct D. In practice, the measurements are very noisy. Thus, matrix D obtained by directly solving (8) for $k = 1, 2, \ldots, 6$ may not be positive semi-definite and is error-prone. It is thus often helpful to take more than six measurements and to use some least squares methods or regularization to obtain a robust estimate while preserving the positive semi-definiteness for physical correctness.

In [60] Wang et al. and in [25] Christiansen et al. proposed an extension of the ROF to denoise tensor-valued data. Two major differences between the two works are that the former regularizes the Cholesky factor of D and uses a channel-by-channel TV regularization, whereas the latter regularizes the tensor D directly and uses a multichannel TV.

The method in [25] is two staged. The first stage is to estimate the diffusion tensors from the raw data based on the Stejskal-Tanner model (8). The obtained

tensors are often noisy and may not be positive semi-definite. The next stage is to use the ROF model to denoise the tensor while restricting the results to be positive semi-definite. The trick they used to ensure positive semi-definiteness is very simple and practical. They observed that a symmetric matrix is positive semi-definite if and only if it has a Cholesky factorization of the form

$$D = LL^T,$$

where L is a lower triangular matrix

$$L = \begin{bmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{bmatrix}.$$

Then one can easily express D in terms of l_{ij} for $1 \leq j \leq i \leq 3$:

$$D = D(L) = \begin{bmatrix} l_{11}^2 & l_{11}l_{21} & l_{11}l_{31} \\ l_{11}l_{21} & l_{21}^2 + l_{22}^2 & l_{21}l_{31} + l_{22}l_{32} \\ l_{11}l_{31} & l_{21}l_{31} + l_{22}l_{32} & l_{31}^2 + l_{32}^2 + l_{33}^2 \end{bmatrix}.$$

The ROF problem, written in a continuous domain, is then formulated as

$$\min_L \left\{ \frac{1}{2} \sum_{ij} \int_\Omega \left[d_{ij}(L) - \hat{d}_{ij} \right]^2 d\mathbf{x} + \lambda \sqrt{\sum_{ij} \left[\int_\Omega \left| \nabla d_{ij}(L) \right| \right]^2} \right\},$$

where $\hat{D} = (\hat{d}_{ij})$ is the observed noisy tensor field, and L is the unknown lower triangular matrix-valued function from Ω to $\mathcal{R}^{3\times 3}$. Here, the matrix-valued version of TV is used. The objective is then differentiated w.r.t. the lower triangular part of L to obtain a system of six first-order optimality conditions. Once the optimal L is obtained, the tensor D can be formed by taking $D = LL^T$ which is a positive semi-definite.

The original ROF problem is strictly convex so that one can obtain the globally optimal solution. However, in this problem, due to the nonlinear change of variables from D to L, the problem becomes non-convex. But the authors of [25] reported that in their experiments, different initial data often resulted in the same solution, so that the non-convexity does not pose any significant difficulty to the optimization of the objective.

4 Numerical Methods and Case Examples

Fast numerical methods for TV minimization continue to be an active research area. Researchers from different fields have been bringing many fresh ideas to the

problem and led to many exciting results. Some categories of particular mention are dual/primal-dual methods, Bregman iterative methods, and graph cut methods. Many of these methods have a long history with a great deal of general theories developed. But when it comes to their application to the ROF model, many further properties and specialized refinements can be exploited to obtain even faster methods. Having said so, different algorithms may adopt different versions of TV. They have different properties and thus may be used for different purposes. Thus, some caution needs to be taken when one attempts to draw conclusions such as method A is faster than method B. Moreover, different methods have different degree of generality. Some methods can be extended directly to deblurring, while some can only be applied to denoising. (Of course, one can use an outer iteration to solve a deblurring problem by a sequence of denoising problems, so that any denoising algorithm can be used. But the convergence of the outer iteration has little, if not none, to do with the inner denoising algorithm.) This section surveys some recent methods for TV denoising and/or deblurring. The model considered here is a generalized ROF model which simultaneously performs denoising and deblurring. The objective function reads

$$F(u) = \frac{1}{2}\int_{\Omega} (Ku - f)^2 \, d\mathbf{x} + \lambda \int_{\Omega} |\nabla u|, \tag{9}$$

where K is a blurring operator and $\lambda > 0$ is the regularization parameter. For simplicity, we assume that K is invertible. When K is the identity operator, (9) is the ROF denoising model.

Dual and Primal-Dual Methods

The ROF objective is non-differentiable in flat regions where $|\nabla u| = 0$. This leads to much difficulty in the optimization process since gradient information (hence, Taylor's expansion) becomes unreliable in predicting the function value even locally. Indeed, the staircase effects of TV minimization can introduce some flat regions which make the problem worse. Even if the standard procedure of replacing the TV with a reasonably smoothed version is used so that the objective becomes differentiable, the Euler-Lagrange equation for (9) is still very stiff to solve. Higher-order methods such as Newton's methods often fail to work because higher-order derivatives are even less reliable.

Due to the difficulty in optimizing the ROF objective directly, much recent research has been directed toward solving some reformulated versions. In particular, methods based on dual and primal-dual formulations have been shown to be very fast in practice. Actually, the dual problem (see (12) below) also has its own numerical difficulties to face, e.g., the objective is rank deficient and some extra work is needed to deal with the constraints. But the dual formulation brings many well-developed ideas and techniques from numerical optimization to bear on this problem. Primal-dual methods have also been studied to combine information from

the primal and dual solutions. Several successful dual and primal-dual methods are reviewed.

Chan-Golub-Mulet's Primal-Dual Method

Some early work in dual and primal-dual methods for the ROF model can be found in [13, 20]. In particular, Chan, Golub, and Mulet (CGM) [20] introduced a primal-dual system involving a primal variable u and a Fenchel dual variable $\mathbf{p}$. It remains one of the most efficient methods today and is perhaps the most intuitive one. It is worthwhile to review it and see how it relates to the more recent methods. Their idea is to start with the Euler-Lagrange equation of (9):

$$K^T Ku - K^T f - \lambda \text{div}\left(\frac{\nabla u}{\sqrt{|\nabla u|^2 + \varepsilon}}\right) = 0. \tag{10}$$

Owing to the singularity of the third term, they introduced an auxiliary variable

$$\mathbf{p} = \frac{\nabla u}{\sqrt{|\nabla u|^2 + \varepsilon}}$$

to form the system

$$\mathbf{p}\sqrt{|\nabla u|^2 + \varepsilon} = \nabla u$$
$$K^T Ku - K^T f - \lambda \text{div}\mathbf{p} = 0.$$

Thus, the blowup singularity is canceled. They proposed to solve this system by Newton's method which is well known to converge quadratically locally if the Jacobian of the system is Lipschitz. Global convergence is observed when coupled with a simple Armijo line search [8]. The variable $\mathbf{p}$ is indeed the same as the Fenchel dual variable $\mathbf{g}$ in (1) when $\nabla u \neq 0$ and $\varepsilon = 0$. Thus, $\mathbf{p}$ is a smoothed version of the dual variable $\mathbf{g}$. Without the introduction of the dual variable, a direct application of the Newton's method to the Euler-Lagrange equation (10) often fails to converge because of the small domain of convergence.

Chambolle's Dual Method

A pure dual method is proposed by Chambolle in [14], where the ROF objective is written solely in terms of the dual variable. By the definition of TV in (1), it can be deduced using duality theory that

$$\inf_u \left\{ \frac{1}{2} \int_\Omega (Ku - f)^2 \, d\mathbf{x} + \lambda \int_\Omega |\nabla u| \right\}$$
$$\Longleftrightarrow \inf_u \sup_{|\mathbf{p}| \leq 1} \left\{ \frac{1}{2} \int_\Omega (Ku - f)^2 \, d\mathbf{x} + \lambda \int_\Omega u \text{div}\mathbf{p} \, d\mathbf{x} \right\} \tag{11}$$

$$\Longleftrightarrow \sup_{|\mathbf{p}|\leq 1} \inf_{u} \left\{ \frac{1}{2} \int_{\Omega} (Ku - f)^2 \, d\mathbf{x} + \lambda \int_{\Omega} u \mathrm{div}\mathbf{p} d\mathbf{x} \right\}$$

$$\Longleftrightarrow \sup_{|\mathbf{p}|\leq 1} \left\{ -\frac{\lambda^2}{2} \int_{\Omega} \left| K^{-T} \mathrm{div}\mathbf{p} - \frac{f}{\lambda} \right|^2 d\mathbf{x} \right\} . \tag{12}$$

The resulting problem has a quadratic objective with quadratic constraints. In contrast, the primal objective is only piecewise smooth which is badly behaved when $\nabla u = 0$. Thus the dual objective function is very simple, but additional efforts are needed to handle the constraints. One can write down the Karush-Kuhn-Tucker (KKT) optimality system [8] of the discretized objective, which amounts to solving a nonlinear system of equations involving complementarity conditions and inequality constraints on the Lagrange multipliers. Interestingly, the Lagrange multipliers have a closed-form solution which greatly simplifies the problem. More precisely, the KKT system consists of the equations

$$\mu p = H(\mathbf{p}) \tag{13}$$

$$\mu \left(|\mathbf{p}|^2 - 1 \right) = 0 \tag{14}$$

$$\mu \geq 0 \tag{15}$$

$$|\mathbf{p}|^2 \leq 1, \tag{16}$$

where μ is the nonnegative Lagrange multiplier and

$$H(\mathbf{p}) := \nabla \left[\left(K^T K \right)^{-1} \mathrm{div}\mathbf{p} - \frac{1}{\lambda} K^{-1} f \right].$$

Since

$$\mu |\mathbf{p}| = |H(\mathbf{p})| ,$$

if $|\mathbf{p}| = 1$, then $\mu = |H(\mathbf{p})|$; if $|\mathbf{p}| < 1$, then the complementarity (14) implies $\mu = 0$ and by (13) $H(\mathbf{p}) = 0$, so that μ is also equal to $|H(\mathbf{p})|$. This simplifies the KKT system into a nonlinear system of $\mathbf{p}$ only

$$|H(\mathbf{p})| \, \mathbf{p} = H(\mathbf{p}) .$$

Chambolle proposes a simple semi-implicit scheme to solve the system:

$$\mathbf{p}^{n+1} = \frac{\mathbf{p}^n + \tau H(\mathbf{p}^n)}{\mathbf{p}^n + \tau |H(\mathbf{p}^n)|}.$$

Here, τ is a positive parameter controlling the stepsize. The method is proven to be convergent for any

$$\tau \leq \frac{1}{8} \left\| \left(K^T K \right)^{-1} \right\| , \tag{17}$$

where $||\cdot||$ is the spectral norm. This method is also faithful to the original ROF problem; it does not require approximating the TV by smoothing.

The convergence rate of this method is at most linear, but for denoising problems, it usually converges fast (measured by the relative residual norm of the optimality condition) in the beginning but stagnates after some iterations (at a level several orders of magnitude higher than the machine epsilon). This is very typical for simple relaxation methods. Fortunately, visually good results (measured by the number of pixels having a gray level different from the optimal one after they are quantized to their 8-bit representation) are often achieved before the method stagnates [64]. However, when applied to deblurring, K is usually ill conditioned, so that the stepsize restriction (17) is too stringent. In this case, another outer iteration is often used in conjunction with the method; see the splitting methods in section "Splitting Methods."

Chambolle's method has been successfully adapted to solve a variety of related image processing problems, e.g., the ROF with nonlocal TV [9], multichannel TV [10], and segmentation problems [4]. We remark that many other approaches for solving (12) have been proposed. A discussion of some first-order methods including projected gradient methods and Nesterov methods can be found in [3, 26, 61].

Primal-Dual Hybrid Gradient Method

As mentioned in the beginning of Sect. 4, the primal and dual problems have their own advantages and numerical difficulties to face. It is therefore tempting to combine the best of both. In [64], Zhu and Chan proposed the *primal-dual hybrid gradient* (PDHG) algorithm which alternates between primal and dual formulations.

The method is based on the primal-dual formulation

$$G(u,\mathbf{p}) := \frac{1}{2}\int_\Omega (Ku-f)^2\,d\mathbf{x} + \lambda\int_\Omega u\mathrm{div}\mathbf{p}\,d\mathbf{x} \to \inf_u \sup_{|\mathbf{p}|\le 1};$$

cf. formulation (11). By fixing its two variables one at a time, this saddle point formulation has two subproblems:

$$\sup_{|\mathbf{p}|\le 1} G(u,\mathbf{p}) \quad \text{and} \quad \inf_u G(u,\mathbf{p}).$$

While one may obtain an optimal solution by solving the two subproblems to a high accuracy alternatively, the PDHG method applies only one step of gradient descent/ascent to each of the two subproblems alternatively. The rationale is that when neither of the two variables are optimal, there is little to gain by iterating each subproblem until convergence. Starting with an initial guess u^0, the following two steps are repeated:

$$\begin{aligned}\mathbf{p}^{k+1} &= P_{|\mathbf{p}|\le 1}\left(\mathbf{p}^k - \tau_k\nabla u^k\right)\\ u^{k+1} &= u^k - \theta_k\left(K^T\left(Ku^k - f\right) + \lambda\mathrm{div}\mathbf{p}^{k+1}\right)\end{aligned}$$

Here, $P_{|\mathbf{p}|\leq 1}$ is the projector onto the feasible set $\{\mathbf{p} : |\mathbf{p}| \leq 1\}$. The stepsizes τ_k and θ_k can be chosen to optimize the performance. Some stepping strategies were presented in [64]. In [65], Zhu, Wright, and Chan studied a variety of stepping strategies for a related dual method.

Numerical results in [64] show that this simple algorithm is faster than the split Bregman iteration (see section "Split Bregman Iteration"), which is faster than Chambolle's semi-implicit dual method (see section "Chambolle's Dual Method"). Some interesting connections between the PDHG algorithm and other algorithms such as *proximal forward-backward splitting, alternating minimization, alternating direction method of multipliers, Douglas-Rachford splitting, split inexact Uzawa*, and *averaged gradient* methods applied to different formulations of the ROF model are studied by Esser et al. in [29]. Such connections reveal some convergence theory of the PDHG algorithm in several important cases (special choices of the stepsizes) in a more general setting.

Semi-smooth Newton's Method

Given the dual problem, it is natural to consider other methods to solve its optimality conditions (13)–(16). A standard technique in optimization to handle complementarity and Lagrange multipliers is to combine them into a single equality constraint. Observe that the constraints $a \geq 0$, $b \geq 0$ and $ab = 0$ can be consolidated into the equality constraint

$$\phi(a, b) := \sqrt{a^2 + b^2} - a - b = 0, \tag{18}$$

where ϕ is known as the Fischer-Burmeister function. Therefore, the KKT system (13)–(16) can be written as

$$\begin{aligned} \mu\mathbf{p} &= H(\mathbf{p}) \\ \phi\left(\mu, 1 - |\mathbf{p}|^2\right) &= 0. \end{aligned}$$

Ng et al. [48] observed that this system is semi-smooth and therefore proposed solving this system using a *semi-smooth Newton's method*. In this method, if the Jacobian of the system is not defined in the classical sense due to the system's lack of enough smoothness, then the Jacobian is replaced by a generalized Jacobian evaluated at a nearby point. It is proven that this method converges superlinearly if the system to solve is at least semi-smooth and if the generalized Jacobians at convergence satisfy some invertibility conditions. For the dual problem (12), the Newton's equation may be singular. This problem is fixed by regularizing the Jacobian.

Primal-Dual Active-Set Method

Hintermüller and Kunisch [37] considered the Fenchel dual approach to formulate a constrained quadratic dual problem and derived a very effective active-set method to handle the constraints. The method separates the variables into active and inactive

sets, so that they can be treated differently accordingly to their characteristics. They considered the case of anisotropic discrete TV norm (3), so that the dual variable is bilaterally constrained, i.e., $-1 \leq \mathbf{p} \leq 1$, whereas the constraints in (12) are quadratic. In this setting, superlinear convergence can be established.

To deal with the bilateral constraints on $\mathbf{p}$, they proposed to use the *primal-dual active-set* (PDAS) algorithm. Consider the general quadratic problem,

$$\min_{y, y \leq \psi} \frac{1}{2} \langle y, Ay \rangle - \langle f, y \rangle ,$$

where ψ is a given vector in $\mathcal{R}^n$. This problem includes (12) as a special instance. The KKT conditions are given by

$$\begin{aligned} Ay + \nu &= f, \\ \nu \odot (\psi - y) &= 0, \\ \nu &\geq 0, \\ \psi - y &\geq 0, \end{aligned}$$

where ν is a vector of Lagrange multipliers and $\odot$ denotes the entrywise product. The idea of the PDAS algorithm is to predict the active variables $\mathcal{A}$ and inactive variables $\mathcal{I}$ to speed up the determination of the final active and inactive variables. The prediction is done by comparing the closeness of ν and $\psi - y$ to zero. If $\psi - y$ is c times closer to zero than ν does, then the variable is predicted as active. The PDAS algorithm is given by

1. Initialize y^0, ν^0. Set $k = 0$.
2. Set $\mathcal{I}^k = \{i : \nu_i^k - c(\psi - y^k)_i \leq 0\}$ and $\mathcal{A}^k = \{i : \nu_i^k - c(\psi - y^k)_i > 0\}$.
3. Solve

$$\begin{aligned} Ay^{k+1} + \nu^{k+1} &= f, \\ y^{k+1} &= \psi \quad \text{on} \quad \mathcal{A}^k, \\ \nu^{k+1} &= 0 \quad \text{on} \quad \mathcal{I}^k. \end{aligned}$$

4. Stop or set $k = k + 1$ and return to Step 2.

Notice that the constraints $a \geq 0$, $b \geq 0$, and $ab = 0$ can be combined as a single equality constraint:

$$\min(a, cb) = 0$$

for any positive constant c. Thus, the KKT system can be written as

$$\begin{aligned} Ay + \nu &= f, \\ C(y, \nu) &= 0, \end{aligned}$$

where $C(y, \nu) = \min(\nu, c(\psi - y))$ for an arbitrary positive constant c. The function C is piecewise linear, whereas the Fisher-Burmeister formulation (18)

is nonlinear. More importantly, applying Newton's method (using a generalized derivative) to such a KKT system yields exactly the PDAS algorithm. This allows Hintermüller et al. to explain the local superlinear convergence of the PDAS algorithm for a class of optimization problems that include the dual of the anisotropic TV deblurring problem [36]. In [37], some conditional global convergence results based on the properties of the blurring matrix K have also been derived. Their formulation is based on the anisotropic TV norm, and the dual problem requires an extra l^2 regularization term when a deblurring problem is solved.

The dual problem (12) is rank deficient and does not have a unique solution in general. In [37], Hintermüller and Kunisch proposed to add a regularization term, so that the solution is unique. The regularized objective function is

$$\int_\Omega \left| K^{-1}\mathrm{div}\mathbf{p} - \lambda^{-1} f \right|^2 d\mathbf{x} + \gamma \int_\Omega |P\mathbf{p}|^2 \, d\mathbf{x},$$

where P is the orthogonal projector onto the null space of the divergence operator div. Later in [38], Hintermüller and Stadler showed that adding such a regularization term to the dual objective is equivalent to smoothing out the singularity of the TV in the primal objective. More precisely, the smoothed TV is given by $\int_\Omega \Phi(|\nabla f|)\, d\ \mathbf{x}$, where

$$\Phi(s) = \begin{cases} s & \text{if } |s| \geq \gamma, \\ \frac{\gamma}{2} + \frac{1}{2\gamma} s^2 & \text{if } |s| < \gamma. \end{cases}$$

An advantage of using this smoothed TV is that the staircase artifacts are reduced.

In [41, 42], Krishnan et al. considered the TV deblurring problem with bound constraints on the image u. An algorithm, called *nonnegatively constrained CGM*, combining the CGM and the PDAS algorithms has been proposed. The image u and its dual $\mathbf{p}$ are treated as in the CGM method, whereas the bound constraints on u are treated as in the PDAS method. The resulting optimality conditions are shown to be semi-smooth. The scheme can also be interpreted as a *semi-smooth quasi-Newton's method* and is proven to converge superlinearly. The method is formulated for isotropic TV, but it can also be applied to anisotropic TV after minor changes.

However, Hintermüller and Kunisch's PDAS method [37] can only be applied to anisotropic TV because they used PDAS that can only handle linear constraints to treat the constraints on $\mathbf{p}$.

Bregman Iteration

Original Bregman Iteration

The *Bregman iteration* is proposed by Osher et al. in [49] for TV denoising. It has also been generalized to solving many convex inverse problems, e.g., [12]. In each

step, the signal removed in the previous step is added back. This is shown to alleviate the loss of contrast problem presented in the ROF model. Starting with the noisy image $f_0 = f$, the following steps are repeated for $j = 0, 1, 2, \ldots$:

1. Set

$$u_{j+1} = \arg\min_u \left\{ \frac{1}{2} \int_\Omega \left(u - f_j\right)^2 d\mathbf{x} + \lambda \int_\Omega |\nabla u| \right\}.$$

2. Set $f_{j+1} = f_j + (f - u_{j+1})$.

In the particular case when f consists of a disk over a constant background, it can be proved that the loss of contrast can be totally recovered. Some theoretical analysis of the method can be found in [49].

For a general regularization functional $J(u)$, the Bregman distance is defined as

$$D_J^p(u, v) = J(u) - J(v) - \langle p, u - v \rangle,$$

where p is an element of the subgradient of J. In case of TV denoising, $J(u) = \lambda \int_\Omega |\nabla u|$. Then, starting with $f_0 = f$, the Bregman iteration is given by

1. Set

$$u_{j+1} = \arg\min_u \left\{ \frac{1}{2} \int_\Omega (u - f)^2 d\mathbf{x} + D_J^{p_j}(u, u_j) \right\}.$$

2. Set $f_{j+1} = f_j + (f - u_{j+1})$.
3. Set $p_{j+1} = f_{j+1} - f$.

In fact, steps 2 and 3 can be combined to $p_{j+1} = p_j + f - u_{j+1}$ without the need of keeping track of f_j. The above expression is for illustrating how the residual is added back to f_j. In this iteration, it has been shown that the Bregman distance between u_j and the *clean image* is monotonically decreasing as long as the L_2-distance is larger than the magnitude of the noise component. But if one iterates until convergence, then $u_j \to f$, i.e., one just gets the noisy image back. This counterintuitive feature is indeed essential to solving other TV minimization problems, e.g., the basis pursuit problem presented next.

The Basis Pursuit Problem

An interesting feature of the Bregman iteration is that, in the discrete setting, if one replaces the term $||u - f||^2$ in the objective by $||Au - f||^2$, where $Au = f$ is underdetermined, then upon convergence of the Bregman iterations, one obtains the solution of the following *basis pursuit problem* [63]:

$$\min_u \{J(u) \,|\, Au = f\}.$$

When $||Au-f||^2$ is used in the objective instead of $||u-f||^2$, the Bregman iteration is given by:

1. Set

$$u_{j+1} = \arg\min_u \left\{ \frac{1}{2} \int_\Omega (Au - f)^2 \, d\mathbf{x} + D_J^{p_j} \left(u, u_j\right) \right\}.$$

2. Set $f_{j+1} = f_j + (f - Au_{j+1})$.
3. Set $p_{j+1} = A^T \left(f_{j+1} - f\right)$.

Split Bregman Iteration

Recently, Goldstein and Osher [35] proposed the *split Bregman iteration* which can be applied to solve the ROF problem efficiently. The main idea is to introduce a new variable so that the TV minimization becomes an L^1 minimization problem which can be solved efficiently by the Bregman iteration. This departs from the original Bregman iteration which solves a sequence of ROF problems to improve the quality of the restored image by bringing back the loss signal. The original Bregman iteration is not iterated until convergence. Moreover, it assumes the availability of a basic ROF solver. The split Bregman method, on the other hand, is an iterative method whose iterates converge to the solution of the ROF problem. In this method, a new variable $\mathbf{q} = \nabla u$ is introduced into the objective function:

$$\min_{u,\mathbf{q}} \left\{ \frac{1}{2} \int_\Omega (u - f)^2 \, d\mathbf{x} + \lambda \int_\Omega |\mathbf{q}| \, d\mathbf{x} \right\}. \tag{19}$$

This problem is solved using a penalty method to enforce the constraint $\mathbf{q} = \nabla u$. The objective with an added penalty is given by

$$G\left(u, \mathbf{q}\right) = \frac{\alpha}{2} \int_\Omega |\mathbf{q} - \nabla u|^2 \, d\mathbf{x} + \frac{1}{2} \int_\Omega (u - f)^2 \, d\mathbf{x} + \lambda \int_\Omega |\mathbf{q}| \, d\mathbf{x}. \tag{20}$$

Notice that if the variables $(u,\ \mathbf{q})$ are denoted by $\mathbf{y}$, then the above objective can be identified as

$$\min_{\mathbf{y}} \left\{ \frac{\alpha}{2} \int_\Omega |A\mathbf{y}|^2 \, d\mathbf{x} + J\left(\mathbf{y}\right) \right\},$$

where

$$\begin{aligned} A\mathbf{y} &= \mathbf{q} - \nabla u, \\ J\left(\mathbf{y}\right) &= \tfrac{1}{2} \textstyle\int_\Omega (u - f)^2 \, d\mathbf{x} + \lambda \int_\Omega |\mathbf{q}| \, d\mathbf{x} \end{aligned}.$$

This is exactly the basis pursuit problem when $\alpha \to \infty$. Actually, even with a fixed finite α, as mentioned in section "The Basis Pursuit Problem," when the Bregman

iteration is used, it converges to the solution of the problem

$$\min_{\mathbf{y}} \{J(\mathbf{y}) \,|\, A\mathbf{y} = 0\},$$

so that the constraint $\mathbf{q} = \nabla u$ is satisfied at convergence.

It is interesting to note that the split Bregman iteration can be viewed as a forward-backward splitting method [53]. Yet another point of view is provided next.

Augmented Lagrangian Method

In [62, 63], it is recognized that the split Bregman iteration is an *augmented Lagrangian method* [33]. This explains some good convergence behaviour of the split Bregman iteration. To motivate the augmented Lagrangian method, consider a general objective function $J(u)$ with equality constraint $H(u) = 0$. The idea of penalty methods is to solve a sequence of unconstrained problems

$$\min_{u} \left\{ J(u) + \frac{1}{\beta} \|H(u)\|^2 \right\}$$

with $\beta \to 0^+$, so that the constraint $H(u) = 0$ is enforced asymptotically. However, one may run into the embarrassing situation where both $H(u(\beta))$ (where $u(\beta)$ is the optimal u for a given β) and β converge to zero in the limit. This could mean that the objective function is stiff when β is very small. The idea of augmented Lagrangian methods is to use a *fixed* parameter. But the penalty term is added to the Lagrangian function, so that the resulting problem is equivalent to the original problem even without letting $\beta \to 0^+$. The augmented Lagrangian function is

$$L(u, v) = J(u) + v \cdot H(u) + \frac{1}{\beta} \|H(u)\|^2,$$

where v is a vector of Lagrange multipliers. Solving $\frac{\partial L}{\partial u} = \frac{\partial L}{\partial v} = 0$ for a saddle point yields exactly $H(u) = 0$ for any $\beta > 0$. The Bregman iteration applied to the penalized objective (20) is indeed computing a saddle point of the augmented Lagrangian function of (19) rather than optimizing (20) itself. Therefore, the constraint $\nabla u = \mathbf{q}$ accompanied with (19) is exact even with a fixed α.

Graph Cut Methods

Recently, there is a burst of interest in graph cut methods for solving various variational problems. The promises of these methods are that they are fast for many practical problems and they can provide globally optimal solution even for "non-convex problems." The discussion below is extracted from [15, 27]. Readers are referred to [15, 27] and the references therein for a more thorough discussion of the subject. Since graph cut problems are combinatoric, the objective has to be cast in

a fully discrete way. That is, not only the image domain has to be discretized to a finite set but also the range of the intensity values. Therefore, in this framework, the given m-by-n image f is a function from $\mathbb{Z}_m \times \mathbb{Z}_n$ to $\mathbb{Z}_K$. The ROF problem thus becomes

$$F(u) = \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n}\left(u_{i,j,} - f_{i,j}\right)^2 + \lambda \|u\|_{TV} \to \min_{u:\mathbb{Z}_m\times\mathbb{Z}_n\to\mathbb{Z}_K},$$

where $||u||_{TV}$ is a discrete TV (4). The next question is how to transform this problem to a graph cut problem in such a way that it can be solved efficiently. It turns out that the (fully discretized) ROF problem can be converted to a finite sequence of graph cut problems. This is due to the co-area formula which is unique to TV. Details are described next.

Leveling the Objective

Some notations and basic concepts are in place. For simplicity, the following discrete TV is adopted:

$$\|u\|_{TV} = \sum_{i=1}^{m-1}\sum_{j=1}^{n-1}\left|u_{i+1,j} - u_{i,j}\right| + \left|u_{i,j+1} - u_{i,j}\right|,$$

which is the anisotropic TV in (3), but with the range of u restricted to $\mathbb{Z}_K$. Recall that the binary image u^k is defined such that each $u^k_{i,j}$ equals 1 if $u_{i,j} \le k$ and equals 0 otherwise. Thus, it is the kth lower level set of u. Then the co-area formula states that the discrete TV can be written as

$$\|u\|_{TV} = \sum_{k=0}^{K-2}\left\|u^k\right\|_{TV}.$$

Thus, it reduces to the TV of each "layer". Note that the TV of the $(K-1)$st level set must be zero, and therefore the above sum is only up to $K-2$.

The fitting term in the objective can also be treated similarly as follows. Notice that for any function $g_{i,j}(s)$, it holds that

$$\begin{aligned} g_{i,j}(s) &= \sum_{k=0}^{s-1}\left[g_{i,j}(k+1) - g_{i,j}(k)\right] + g_{i,j}(0) \\ &= \sum_{k=0}^{K-2}\left[g_{i,j}(k+1) - g_{i,j}(k)\right]\chi_{k<s} + g_{i,j}(0), \end{aligned}$$

where $\chi_{k<s} = 1$ if $k < s$ and 0 otherwise. Define $g_{i,j}(k) = \frac{1}{2}(s - f_{i,j})^2$. Then,

$$\begin{aligned}\tfrac{1}{2}\left(u_{i,j} - f_{i,j}\right)^2 &= g_{i,j}\left(u_{i,j}\right) \\ &= \sum_{k=0}^{K-2}\left[g_{i,j}\left(k+1\right) - g_{i,j}\left(k\right)\right]\chi_{k<u_{i,j}} + g_{i,j}\left(0\right) \\ &= \sum_{k=0}^{K-2}\left[g_{i,j}\left(k+1\right) - g_{i,j}\left(k\right)\right]\left(1 - u_{i,j}^k\right) + g_{i,j}\left(0\right).\end{aligned}$$

As a result, the ROF objective can be expressed as

$$\sum_{k=0}^{K-2}\left\{\sum_{i,j}\left[g_{i,j}\left(k+1\right) - g_{i,j}\left(k\right)\right]\left(1 - u_{i,j}^k\right) + \lambda\left\|u^k\right\|_{TV}\right\} + C,$$

where $C = \sum_{i,j} g_{i,j}\left(0\right)$.

By defining the objective function

$$F^k\left(v^k\right) = \sum_{i,j}\left[g_{i,j}\left(k+1\right) - g_{i,j}\left(k\right)\right]\left(1 - v_{i,j}^k\right) + \lambda\left\|v^k\right\|_{TV},$$

where v^k is a binary function, the ROF problem is seen to be equivalent to

$$\min_{v^1, v^2, \ldots, v^{K-2}} \sum_{k=0}^{K-2} F^k\left(v^k\right)$$

subject to the inclusion constraints $v_{i,j}^k \leq v_{i,j}^{k+1}$ for all i, j, k. The constraints make sure the binary functions $\{v^k\}_k$ define the lower level sets of some function v. A very important result is that the minimization can be done independently for each v^k; amazingly, the solutions $\{v^k\}$ satisfy the inclusion property automatically! See [27] for further details.

Defining a Graph

To minimize each F^k w.r.t. a binary function v^k, a graph cut method is used. First observe that since $g_{i,j}(k) = \frac{1}{2}(k - f_{i,j})^2$, F^k can be simplified to

$$F^k\left(v^k\right) = \sum_{i,j}\left[\frac{1}{2} + \left(k - f_{i,j}\right)\right]\left(1 - v_{i,j}^k\right) + \lambda\left\|v^k\right\|_{TV}.$$

By absorbing some constants and dropping the superscript on v^k, the objective takes the following form:

$$F^k\left(v\right) = \sum_{i,j}\left(f_{i,j} - k - \frac{1}{2}\right)v_{i,j} + \lambda\left\|v\right\|_{TV}. \tag{21}$$

Then, a graph with $mn + 2$ nodes is constructed in the following way:

1. Each of the mn pixels is a node, labeled by (i, j) for $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, n$.
2. Add two additional nodes, called the source S and the sink T.
3. For each (i, j), connect it to $(i \pm 1, j)$ and $(i, j \pm 1)$ with capacity λ.
4. For each (i, j), connect S to it with capacity $\frac{1}{2} + k - f_{i,j}$ if $\frac{1}{2} + k - f_{i,j} > 0$ and connect it to T with capacity $f_{i,j} - k - \frac{1}{2}$ if $f_{i,j} - k - \frac{1}{2} > 0$.

A cut (a.k.a. an *st*-cut) in the graph is a partition $(\mathcal{S}, \mathcal{T})$ such that $\mathcal{S} \in \mathcal{S}$ and $T \in \mathcal{T}$. The cost of the cut $\mathcal{C}(\mathcal{S}, \mathcal{T})$ is defined as the sum of the capacities of all edges from $\mathcal{S}$ to $\mathcal{T}$. For a given cut, let $v_{i,j}$ equals 1 if $(i, j) \in \mathcal{S}$ and equals 0 if $(i, j) \in \mathcal{T}$. Then it can be verified that

$$C(\mathcal{S}, \mathcal{T}) = \sum_{i,j} \max\left\{\frac{1}{2} + k - f_{i,j}, 0\right\} v_{i,j} + \max\left\{f_{i,j} - k - \frac{1}{2}, 0\right\} (1 - v_{i,j}) + \lambda \left\|v^k\right\|_{TV}$$

which is the same as F^k in (21), up to the constant $\Sigma_{i,j} \max\{f_{i,j} - k, 0\}$. Therefore, computing the minimum cut is equivalent to minimizing (21). It is also well known that the minimum cut problem is equivalent to the maximum flow problem.

Recall that there are $K - 1$ graphs to cut. A simple way is to do them one by one using any classical maximum flow algorithm. But one can exploit the inclusion property to reduce the work; for instance, see the divide-and-conquer algorithm proposed in [27].

In graph cut methods, a fundamental question is what kind of optimization problems can be transformed to a graph cut problem. A particularly relevant question is whether a function is levelable, i.e., its minimization can be done by first solving the simpler problem on each of its level set, followed by assembling the resulting level sets. Interestingly, the only levelable convex regularization function (satisfying some very natural and mild conditions) is TV [27]. This indicates that TV is much more than just an ordinary semi-norm.

Quadratic Programming

The discrete anisotropic TV is a piecewise linear function. Fu et al. [30] showed that by introducing some auxiliary variables, one can transform the TV to a linear function but with some additional linear constraints. Together with the fitting term, the problem to solve has a quadratic objective function with linear constraints.

The objective function considered is by Fu et al.

$$F(u) = \frac{1}{2} \|Ku - f\|^2 + \lambda \sum_{i,j} \left|u_{i+1,j} - u_{i,j}\right| + \left|u_{i,j+1} - u_{i,j}\right|,$$

which can also be written as

$$F(u) = \frac{1}{2}\|Ku - f\|^2 + \lambda \|Ru\|_1$$

where R is a $2mn$-by-mn matrix. If the original isotropic TV is used, then it cannot be written in this form.

The trick they used is to let $v = Ru$ and then split it into positive and negative parts: $v^+ = \max(v, 0)$ and $v^- = \max(-v, 0)$. Then, the objective can be written as

$$G\left(u, v^+, v^-\right) = \frac{1}{2}\|Ku - f\|^2 + \lambda \left(1^T v^+ + 1^T v^-\right),$$

which is a quadratic function. But some linear constraints are added:

$$\begin{aligned} Ru &= v^+ - v^-, \\ v^+, v^- &\geq 0. \end{aligned}$$

Now, this problem can be solved by standard primal-dual interior-point methods. Here, "dual" refers to the Lagrange multipliers for the linear constraints. The major steps can be summarized as follows:

1. Write down the KKT system of optimality conditions, which has a form of $f(x, \mu, s) = 0$, where $x \geq 0$ is the variable of the original problem ($x = (u, v^+, v^-)$ in the present case); μ is the Lagrange multipliers for the equality constraints; and $s \geq 0$ is the Lagrange multipliers for the inequality constraints.
2. Relax the complementarity $xs = 0$ (part of $f(x, \mu, s) = 0$) to $xs = v$, where $v > 0$.
3. Solve the relaxed problem $f_v(x, \mu, s) = 0$ by Newton's method.
4. After each Newton's iteration, reduce the value of v so that the solution of $f(x, \mu, s) = 0$ is obtained at convergence.

In this method, the relaxed complementarity $xs = v$ forces the variables x, s to lie in the interior of the feasible region. Once the variables are away from the boundary, the problem becomes a nice unconstrained quadratic problem locally. The main challenge here is that the linear system to solve in each Newton's iteration becomes increasingly ill conditioned. Under this framework, bound constraints such as $u_{\min} \leq u \leq u_{\max}$ or any linear equality constraints can be easily added.

Second-Order Cone Programming

The trick to "linearize" the TV presented in the last section does not work for isotropic TV. Goldfarb and Yin [34] proposed a *second-order cone programming*

(SOCP) formulation which works for the isotropic version (2). Moreover, its connection to SOCP allows the use of available SOCP solvers to obtain the solutions. The problem they considered is the constrained ROF problem:

$$\min_u \|u\|_{TV}$$

subject to

$$\|u - f\| \leq \sigma,$$

where σ is the standard deviation of the noise which is assumed to be known.

Let $w^x_{i,j} = u_{i+1,j} - u_{i,j}$ and $w^y_{i,j} = u_{i,j+1} - u_{i,j}$. The TV becomes

$$\sum_{i,j} \sqrt{\left(w^x_{i,j}\right)^2 + \left(w^y_{i,j}\right)^2}.$$

By introducing the variables $v = f - u$ and t and the constraint

$$\left(w^x_{i,j}\right)^2 + \left(w^y_{i,j}\right)^2 \leq t^2_{i,j},$$

the TV minimization problem becomes

$$\begin{aligned} & \min \sum_{i,j} t_{i,j} \\ s.t. \quad & u + v = f \\ & w^x_{i,j} = u_{i+1,j} - u_{i,j} \\ & w^y_{i,j} = u_{i,j+1} - u_{i,j} \\ & (\sigma, v) \in \text{cone}^{mn+1} \\ & \left(t_{i,j}, w^x_{i,j}, w^y_{i,j}\right) \in \text{cone}^3. \end{aligned}$$

Here, cone ^{n} is the second-order cone in $\mathcal{R}^n$:

$$\{x \in \mathbb{R}^n : \|(x_2, x_3, \ldots, x_n)\| \leq x_1\}.$$

The optimal solution satisfies

$$t^2_{i,j} = \left(w^x_{i,j}\right)^2 + \left(w^y_{i,j}\right)^2,$$

so that

$$\sum_{i,j} t_{i,j} = \sum_{i,j} \sqrt{\left(w^x_{i,j}\right)^2 + \left(w^y_{i,j}\right)^2} = \|u\|_{TV}.$$

A SOCP formulation of the dual ROF problem is also given in [34].

The SOCP can be solved by interior-point methods. The above formulation can be slightly simplified by eliminating u. But the number of variables (hence, the size of the Newton's equation) is still several times larger than the original problem. Goldfarb and Yin proposed a domain decomposition method to split the large programming problem into smaller ones, so that each subproblem can be solved efficiently. Of course, the convergence rate of the method deteriorates as the domain is further split.

Majorization-Minimization

Majorization-minimization(MM) (or *minorization-maximization*) [43] is a well-studied technique in optimization. The main idea is that at each step of the method, the objective function is replaced by a simple one, called the *surrogate function*, such that its minimization is easy to carry out and the result gives a smaller objective value of the original problem. For a given objective, usually many surrogate functions are possible. In many cases, one can even reduce multidimensional problems into a set of one-dimensional problems. Methods of this class have been heavily used in statistics communities. Indeed expectation-maximization (EM) algorithms are special cases of MM.

The use of MM to solving discrete TV problems can be traced back to the study of emission and transmission tomography reconstruction problems by Lange and Carson in [44]. Recently, some authors have applied the method to solving TV deblurring problems [6]. However, the method is actually the same as the classical lagged diffusivity fixed point iteration proposed by [58] for the particular surrogate function used in [6]. Nevertheless, it is still worthy to present the framework here because other surrogate functions can lead to different schemes.

Denote by u^k the kth iterate. In this method, the surrogate function (majorizer) $Q(u|u^k)$ is defined such that

$$\begin{aligned} F\left(u^k\right) &= Q\left(u^k|u^k\right) \\ F\left(u\right) &\leq Q\left(u|u^k\right), \text{for all } u \end{aligned}$$

Then, the next iterate is defined to be the minimizer of the surrogate function

$$u^{k+1} := \arg\min_u Q\left(u|u^k\right).$$

In this way, the following monotonic decreasing property holds

$$F\left(u^{k+1}\right) \leq Q\left(u^{k+1}|u^k\right) \leq Q\left(u^k|u^k\right) = F\left(u^k\right).$$

Presumably, the function Q should be chosen so that its minimum is easy to compute. In many applications, it may even be chosen to have a separable form

$$Q\left(u|u^k\right) = Q_1\left(u_1|u^k\right) + Q_2\left(u_2|u^k\right) + \cdots + Q_n\left(u_n|u^k\right),$$

so that its minimization reduces to n's one-dimensional (1D) problems. A promise of this method is that each iteration is very easy to carry out, which compensates its linear-only convergence. To construct a surrogate Q_{TV} for TV, first note that

$$\sqrt{a} = \left(\sqrt[4]{b}\right)\left(\frac{\sqrt{a}}{\sqrt[4]{b}}\right) \leq \frac{\sqrt{b}}{2} + \frac{a}{2\sqrt{b}}$$

for all $a, b \geq 0$. Let D_x and D_y be the forward difference operator in x and in y directions, respectively. Then,

$$\begin{aligned}\|u\|_{TV} &= \sum_{i,j} \sqrt{\left(D_x u_{i,j}\right)^2 + \left(D_y u_{i,j}\right)^2} \\ &\leq \tfrac{1}{2} \sum_{i,j} \sqrt{\left(D_x u_{i,j}^k\right)^2 + \left(D_y u_{i,j}^k\right)^2} + \tfrac{1}{2} \sum_{i,j} \frac{\left(D_x u_{i,j}\right)^2 + \left(D_y u_{i,j}\right)^2}{\sqrt{\left(D_x u_{i,j}^k\right)^2 + \left(D_y u_{i,j}^k\right)^2}}\end{aligned}$$

The surrogate is thus defined as

$$Q_{TV}\left(u|u^k\right) = \frac{1}{2}\left\|u^k\right\|_{TV} + \frac{1}{2} \sum_{i,j} \frac{\left(D_x u_{i,j}\right)^2 + \left(D_y u_{i,j}\right)^2}{\sqrt{\left(D_x u_{i,j}^k\right)^2 + \left(D_y u_{i,j}^k\right)^2}}$$

which is quadratic in u. Notice that the 2D discrete gradient matrix is given by

$$\nabla = \begin{bmatrix} \nabla_n \otimes I_m \\ I_n \otimes \nabla_m \end{bmatrix},$$

where ∇_m is the m-by-m1D forward difference matrix (under Neumann boundary conditions)

$$\nabla_m = \begin{bmatrix} -1 & 1 & & \\ & -1 & 1 & \\ & & \ddots & \ddots \\ & & & -1 \quad 1 \\ & & & 0 \end{bmatrix}$$

Let $\lambda_{i,j}^k = 1/\sqrt{\left(D_x u_{i,j}^k\right)^2 + \left(D_y u_{i,j}^k\right)^2}$ and let

$$\Lambda^k = \operatorname{diag}\left(\lambda_{1,1}^k, \ldots, \lambda_{m,n}^k, \lambda_{1,1}^k, \ldots, \lambda_{m,n}^k\right).$$

The surrogate becomes

$$Q_{TV}\left(u|u^k\right) = \frac{1}{2}\left\|u^k\right\|_{TV} + \frac{1}{2} u^T \nabla^T \Lambda^k \nabla u.$$

In this case, the minimization of Q_{TV} cannot be reduced to a set of 1D problems. But it does become quadratic.

Finally, the majorizer for the ROF model is

$$Q\left(u|u^k\right) = \frac{1}{2}\|Ku - f\|^2 + \lambda Q_{TV}\left(u|u^k\right).$$

While this method completely bypasses the need to optimize the TV term directly, each iteration requires solving the linear system

$$\left(K^T K + \lambda \nabla^T \Lambda^k \nabla\right) u^{k+1} = K^T f.$$

This scheme is exactly the lagged diffusivity fixed point iteration. Assume that K is full rank, then the linear system is positive definite. A standard way is to use preconditioned conjugate gradient to solve. Many preconditioners have been proposed for this problem in the 1990s, e.g., cosine transform and multigrid and multiplicative operator splitting; see [17] and the references therein. However, due to the highly varying coefficients in Λ^k, it can be nontrivial to solve efficiently.

Splitting Methods

Recently, there have been several proposals for solving TV deblurring problems based on the idea of separating the deblurring process and the TV regularization process. Many of them are based on the idea that the minimization of an objective of the form

$$F(u) = J_1(u) + J_2(Au),$$

with Aa linear operator, can be approximated by the minimization of either of the following two objectives:

$$G(u, v) = J_1(u) + \tfrac{\alpha}{2}\|u - v\|^2 + J_2(Av),$$
$$G(u, v) = J_1(u) + \tfrac{\alpha}{2}\|Au - v\|^2 + J_2(v),$$

where α is a large scalar. Then G is minimized w.r.t. u and v alternatively. In this way, at each iteration, the minimization of J_1 and J_2 is done separately. The same idea can be generalized to split an objective with n terms to an objective with n variables.

Consider the discrete ROF model:

$$F(u) = \frac{1}{2}\|Ku - f\|^2 + \lambda \||\nabla u|\|_1.$$

Huang et al. [39] and Bresson and Chan [10] considered the splitting

$$G(u,v) = \frac{1}{2}\|Ku - f\|^2 + \frac{\alpha}{2}\|u - v\|^2 + \lambda \||\nabla v|\|_1 .$$

In this case, the minimization w.r.t. u becomes

$$\left(K^T K + \alpha I\right) u = K^T f + \alpha v,$$

which can be solved with the fast Fourier transform (FFT) in $O(N \log N)$ operations when the blurring matrix K can be diagonalized by a fast transform matrix. The minimization w.r.t. v is the ROF denoising problem which can be solved using any of the aforementioned denoising method. Both [39] and [10] employed Chambolle's dual algorithm. The point is that solving TV denoising is much easier than solving TV deblurring (directly). Moreover, some algorithms such as those based on graph cut cannot be applied to deblurring directly. The reason is that the pixel values in the fitting are no longer separable, which in turn makes the fitting term not "levelable." However, using the splitting technique, one can now apply graph cut methods to solve each denoising problem.

This method is generally very fast. Moreover, it often works for a large range of α. But when α is too large, the Chambolle's iteration may slow down. This splitting method has also been applied to other image processing problems such as segmentation [10].

An alternative splitting is proposed by Wang et al. [59]. The bivariate function they used is given by

$$G(u,v) = \frac{1}{2}\|Ku - f\|^2 + \frac{\alpha}{2}\||\nabla u - v|\|^2 + \lambda \||v|\|_1 .$$

The minimization w.r.t. u requires solving

$$\left(K^T K - \alpha \Delta\right) u = K^T f + \alpha v,$$

where Δ is the 2D Laplacian. This equation can again be solved with FFT in $O(N \log N)$ operations. The minimization w.r.t. $v$gs decoupled into N minimization problems (one for each pixel) of two variables. A simple closed-form solution for the 2D minimization problems is available. Therefore, the computation cost per iteration is even less than the approach taken in [39] and [10]. Remark that this objective is indeed the same as the split Bregman method (20). A difference is that when the split Bregman iteration converges, it holds exactly that $\nabla u = v$. But the simple alternating minimization used in most splitting methods does not guarantee $\nabla u = v$ at convergence.

An alternative splitting is introduced by Bect et al. in [5]. It is based on the observation that, for any symmetric positive definite matrix B with $||B|| < 1$, it holds that

$$\langle Bv, v\rangle = \min_{u \in \mathbb{R}^N} \left\{\|u - v\|^2 + \langle Cu, u\rangle\right\}$$

for all $v \in \mathcal{R}^N$, where $C = B(I - B)^{-1}$. Then, the ROF model can be formulated as the minimization of the following bivariate function:

$$G(u, v) = \frac{1}{2\mu}\left(\|u - v\|^2 + \langle Cu, u\rangle\right) + \frac{1}{2}\left(\|f\|^2 - 2\langle Kv, f\rangle\right) + \lambda \||\nabla v|\|_1,$$

where $\mu > 0$ such that $\mu||K^TK|| < 1$ and $B = \mu K^TK$. The minimization of G w.r.t. u has a closed-form solution $u = (I-B)v = (I-\mu K^TK)v$. The minimization of G w.r.t. v is a TV denoising problem. At convergence, the minimizer of F is exactly recovered. An interesting property of this splitting is that it does not involve any matrix inversion in the alternating minimization of G.

5 Conclusion

In this chapter, some recent developments of numerical methods for TV minimization and their applications are reviewed. The chosen topics only reflect the interest of the authors and are by no means comprehensive. It is also hoped that this chapter can serve as a guide to recent literature on some of these recent developments.

Cross-References

- Compressive Sensing
- Duality and Convex Programming
- Energy Minimization Methods
- Large-Scale Inverse Problems in Imaging
- Mumford and Shah Model and Its Applications to Image Segmentation and Image Restoration
- Numerical Methods and Applications in Total Variation Image Restoration
- Regularization Methods for Ill-Posed Problems
- Total Variation in Imaging
- Variational Methods in Shape Analysis

References

1. Acar, A., Vogel, C.: Analysis of bounded variation penalty methods for ill-posed problems. Inverse Probl. **10**(6), 1217–1229 (1994)
2. Adams, R., Fournier, J.: Sobolev Spaces. Volume 140 of Pure and Applied Mathematics, 2nd edn. Academic, New York (2003)
3. Aujol, J.-F.: Some first-order algorithms for total variation based image restoration. J. Math. Imaging Vis. **34**(3), 307–327 (2009)
4. Aujol, J.-F., Gilboa, G., Chan, T., Osher, S.: Structure-texture image decomposition – modeling, algorithms, and parameter selection. Int. J. Comput. Vis. **67**(1), 111–136 (2006)

5. Bect, J., Blanc-Féraud, L., Aubert, G., Chambolle, A.: A l^1-unified variational framework for image restoration. In: Proceedings of ECCV. Volume 3024 of Lecture Notes in Computer Sciences, Prague, Czech Republic, pp. 1–13 (2004)
6. Bioucas-Dias, J., Figueiredo, M., Nowak, R.: Total variation-based image deconvolution: a majorization-minimization approach. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2006), Toulouse, France, vol. 2, pp. 14–19 (2006)
7. Blomgren, P., Chan, T.: Color TV: total variation methods for restoration of vector-valued images. IEEE Trans. Image Process. **7**, 304–309 (1998)
8. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press, Cambridge (2004)
9. Bresson, X., Chan, T.: Non-local unsupervised variational image segmentation models. UCLA CAM Report, 08–67 (2008)
10. Bresson, X., Chan, T.: Fast dual minimization of the vectorial total variation norm and applications to color image processing. Inverse Probl. Imaging **2**(4), 455–484 (2008)
11. Buades, A., Coll, B., Morel, J.: A review of image denoising algorithms, with a new one. Multiscale Model. Simul. **4**(2), 490–530 (2005)
12. Burger, M., Frick, K., Osher, S., Scherzer, O.: Inverse total variation flow. Multiscale Model. Simul. **6**(2), 366–395 (2007)
13. Carter, J.: Dual methods for total variation-based image restoration. Ph.D. thesis, UCLA, Los Angeles (2001)
14. Chambolle, A.: An algorithm for total variation minimization and applications. J. Math. Imaging Vis. **20**, 89–97 (2004)
15. Chambolle, A., Darbon, J.: On total variation minimization and surface evolution using parametric maximum flows. Int. J. Comput. Vis. **84**(3), 288–307 (1997)
16. Chambolle, A., Lions, P.: Image recovery via total variation minimization and related problems. Numer. Math. **76**, 167–188 (1997)
17. Chan, R., Chan, T., Wong, C.: Cosine transform based preconditioners for total variation deblurring. IEEE Trans. Image Process. **8**, 1472–1478 (1999)
18. Chan, R., Wen, Y., Yip, A.: A fast optimization transfer algorithm for image inpainting in wavelet domains. IEEE Trans. Image Process. **18**(7), 1467–1476 (2009)
19. Chan, T., Vese, L.: Active contours without edges. IEEE Trans. Image Process. **10**(2), 266–277 (2001)
20. Chan, T., Golub, G., Mulet, P.: A nonlinear primal-dual method for total variation-based image restoration. SIAM J. Sci. Comput. **20**, 1964–1977 (1999)
21. Chan, T., Esedoḡlu, S., Park, F., Yip, A.: Recent developments in total variation image restoration. In: Paragios, N., Chen, Y., Faugeras, O. (eds.) Handbook of Mathematical Models in Computer Vision. Springer, Berlin, pp. 17–32 (2005)
22. Chan, T., Esedoḡlu, S., Nikolova, M.: Algorithms for finding global minimizers of image segmentation and denoising models. SIAM J. Appl. Math. **66**(5), 1632–1648 (2006)
23. Chan, T., Shen, J., Zhou, H.: Total variation wavelet inpainting. J. Math. Imaging Vis. **25**(1), 107–125 (2006)
24. Chan, T., Ng, M., Yau, C., Yip, A.: Superresolution image reconstruction using fast inpainting algorithms. Appl. Comput. Harmon. Anal. **23**(1), 3–24 (2007)
25. Christiansen, O., Lee, T., Lie, J., Sinha, U., Chan, T.: Total variation regularization of matrix-valued images. Int. J. Biomed. Imaging **2007**, 27432 (2007)
26. Combettes, P., Wajs, V.: Signal recovery by proximal forward-backward splitting. Multiscale Model. Simul. **4**(4), 1168–1200 (2004)
27. Darbon, J., Sigelle, M.: Image restoration with discrete constrained total variation part I: fast and exact optimization. J. Math. Imaging Vis. **26**, 261–276 (2006)
28. Efros, A., Leung, T.: Texture synthesis by non-parametric sampling. In: Proceedings of the IEEE International Conference on Computer Vision, Corfu, vol. 2, pp. 1033–1038 (1999)
29. Esser, E., Zhang, X., Chan, T.: A general framework for a class of first order primal-dual algorithms for TV minimization. SIAM J. Imaging Sci. **3**(4), 1015–1046 (2010)

30. Fu, H., Ng, M., Nikolova, M., Barlow, J.: Efficient minimization methods of mixed $l2$-$l1$ and $l1$-$l1$ norms for image restoration. SIAM J. Sci. Comput. **27**(6), 1881–1902 (2006)
31. Gilboa, G., Osher, S.: Nonlocal operators with applications to image processing. Multiscale Model. Simul. **7**(3), 1005–1028 (2008)
32. Giusti, E.: Minimal Surfaces and Functions of Bounded Variation. Birkhäuser, Boston (1984)
33. Glowinki, R., Le Tallec, P.: Augmented Lagrangians and Operator-Splitting Methods in Nonlinear Mechanics. SIAM, Philadelphia (1989)
34. Goldfarb, D., Yin, W.: Second-order cone programming methods for total variation based image restoration. SIAM J. Sci. Comput. **27**(2), 622–645 (2005)
35. Goldstein, T., Osher, S.: The split Bregman method for l^1-regularization problems. SIAM J. Imaging Sci. **2**(2), 323–343 (2009)
36. Hintermüller, M., Kunisch, K.: Total bounded variation regularization as a bilaterally constrained optimisation problem. SIAM J. Appl. Math. **64**, 1311–1333 (2004)
37. Hintermüller, M., Stadler, G.: A primal-dual algorithm for TV-based inf-convolution-type image restoration. SIAM J. Sci. Comput. **28**, 1–23 (2006)
38. Hintermüller, M., Ito, K., Kunisch, K.: The primal-dual active set strategy as a semismooth Newton's method. SIAM J. Optim. **13**(3), 865–888 (2003)
39. Huang, Y., Ng, M., Wen, Y.: A fast total variation minimization method for image restoration. Multiscale Model. Simul. **7**(2), 774–795 (2008)
40. Kanwal, R.P.: Generalized Functions: Theory and Applications. Birkhäuser, Boston (2004)
41. Krishnan, D., Lin, P., Yip, A.: A primal-dual active-set method for non-negativity constrained total variation deblurring problems. IEEE Trans. Image Process. **16**(2), 2766–2777 (2007)
42. Krishnan, D., Pham, Q., Yip, A.: A primal dual active set algorithm for bilaterally constrained total variation deblurring and piecewise constant Mumford-Shah segmentation problems. Adv. Comput. Math. **31**(1–3), 237–266 (2009)
43. Lange, K.: (2004) Optimization. Springer, New York
44. Lange, K., Carson, R.: (1984) EM reconstruction algorithms for emission and transmission tomography. J. Comput. Assist. Tomogr. **8**, 306–316
45. Law, Y., Lee, H., Yip, A.: A multi-resolution stochastic level set method for Mumford-Shah image segmentation. IEEE Trans. Image Process. **17**(3), 2289–2300 (2008)
46. LeVeque, R.: Numerical Methods for Conservation Laws, 2nd edn. Birkhäuser, Basel (2005)
47. Mumford, D., Shah, J.: Optimal approximation by piecewise smooth functions and associated variational problems. Commun. Pure Appl. Math. **42**, 577–685 (1989)
48. Ng, M., Qi, L., Tang, Y., Huang, Y.: On semismooth Newton's methods for total variation minimization. J. Math. Imaging Vis. **27**(3), 265–276 (2007)
49. Osher, S., Burger, M., Goldfarb, D., Xu, J., Yin, W.: An iterative regularization method for total variation based image restoration. Multiscale Model. Simul. **4**, 460–489 (2005)
50. Royden, H.: Real Analysis, 3rd edn. Prentice-Hall, Englewood Cliffs (1988)
51. Rudin, L., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. Physica D **60**, 259–268 (1992)
52. Sapiro, G., Ringach, D.: Anisotropic diffusion of multivalued images with applications to color filtering. IEEE Trans. Image Process. **5**, 1582–1586 (1996)
53. Setzer, S.: Split Bregman algorithm, Douglas-Rachford splitting and frame shrinkage. In: Proceedings of Scale-Space, Voss, Norway, pp. 464–476 (2009)
54. Setzer, S., Steidl, G., Popilka, B., Burgeth, B.: Variational methods for denoising matrix fields. In: Laidlaw, D., Weickert, J. (eds.) Visualization and Processing of Tensor Fields: Advances and Perspectives, Mathematics and Visualization, pp. 341–360. Springer, Berlin (2009)
55. Shen, J., Kang, S.: Quantum TV and application in image processing. Inverse Probl. Imaging **1**(3), 557–575 (2007)
56. Strang, G.: Maximal flow through a domain. Math. Program. **26**(2), 123–143 (1983)
57. Tschumperlé, D., Deriche, R.: Diffusion tensor regularization with constraints preservation. In: Proceedings of 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Kauai, vol. 1, pp. 948–953. IEEE Computer Science Press (2001)

58. Vogel, C., Oman, M.: Iteration methods for total variation denoising. SIAM J. Sci. Comput. **17**, 227–238 (1996)
59. Wang, Y, Yang, J., Yin, W., Zhang, Y.: A new alternating minimization algorithm for total variation image reconstruction. SIAM J. Imaging Sci. **1**(3), 248–272 (2008)
60. Wang, Z., Vemuri, B., Chen, Y., Mareci, T.: A constrained variational principle for direct estimation and smoothing of the diffusion tensor field from complex DWI. IEEE Trans. Med. Imaging **23**(8), 930–939 (2004)
61. Weiss, P., Aubert, G., Blanc-Fèraud, L.: Efficient schemes for total variation minimization under constraints in image processing. SIAM J. Sci. Comput. **31**(3), 2047–2080 (2009)
62. Wu, C., Tai, X.C.: Augmented Lagrangian method, dual methods, and split Bregman iteration for ROF, vectorial TV, and high order models. SIAM J. Imaging Sci. **3**(3), 300–339 (2010)
63. Yin, W., Osher, S., Goldfarb, D., Darbon, J.: Bregman iterative algorithms for l^1-minimization with applications to compressed sensing. SIAM J. Imaging Sci. **1**(1), 143–168 (2008)
64. Zhu, M., Chan, T.: An efficient primal-dual hybrid gradient algorithm for total variation image restoration. UCLA CAM Report, 08–34 (2008)
65. Zhu, M., Wright, S.J., Chan, T.F.: Duality-based algorithms for total-variation-regularized image restoration. Comput. Optim. Appl. **47**(3), 377–400 (2010)

Mumford and Shah Model and Its Applications to Image Segmentation and Image Restoration

Leah Bar, Tony F. Chan, Ginmo Chung, Miyoun Jung, Nahum Kiryati, Nir Sochen, and Luminita A. Vese

Contents

1 Introduction 1540
2 Background: The First Variation 1542
Minimizing in u with K Fixed 1542
3 Minimizing in K 1545
4 Mathematical Modeling and Analysis: The Weak Formulation of the Mumford and Shah Functional 1547
5 Numerical Methods: Approximations to the Mumford and Shah Functional 1550
6 Ambrosio and Tortorelli Phase-Field Elliptic Approximations 1551
7 Approximations of the Perimeter by Elliptic Functionals 1551
8 Ambrosio-Tortorelli Approximations 1552
9 Level Set Formulations of the Mumford and Shah Functional 1554
10 Piecewise-Constant Mumford and Shah Segmentation Using Level Sets 1558
11 Piecewise-Smooth Mumford and Shah Segmentation Using Level Sets 1561
12 Case Examples: Variational Image Restoration with Segmentation-Based Regularization 1567

L. Bar (✉)
Department of Mathematics, Tel Aviv University, Minneapolis, MN, USA
e-mail: barleah@hotmail.com

T.F. Chan
Office of the President, Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong
e-mail: tonyfchan@ust.hk

G. Chung
Los Angeles, CA, United States

M. Jung • L.A. Vese
Department of Mathematics, Hankuk University of Foreign Studies, Los Angeles, CA, USA
e-mail: mjung@hufs.ac.kr

N. Kiryati • N. Sochen
Tel Aviv University, Tel Aviv, Israel
e-mail: nk@eng.tau.ac.il

© Springer Science+Business Media New York 2015

O. Scherzer (ed.), *Handbook of Mathematical Methods in Imaging*,
DOI 10.1007/978-1-4939-0790-8_25

13 Non-blind Restoration 1570
14 Semi-blind Restoration 1570
15 Image Restoration with Impulsive Noise 1573
16 Color Image Restoration 1577
17 Space-Variant Restoration 1579
18 Level Set Formulations for Joint Restoration and Segmentation 1582
19 Image Restoration by Nonlocal Mumford-Shah Regularizers 1584
20 Conclusion 1592
21 Recommended Reading 1593
Cross-References 1594
References 1594

Abstract

This chapter presents an overview of the Mumford and Shah model for image segmentation. It discusses its various formulations, some of its properties, the mathematical framework, and several approximations. It also presents numerical algorithms and segmentation results using the Ambrosio-Tortorelli phase-field approximations on one hand and level set formulations on the other hand. Several applications of the Mumford-Shah problem to image restoration are also presented.

1 Introduction

An important problem in image analysis and computer vision is the segmentation one, aiming to partition a given image into its constituent objects, or to find boundaries of such objects. This chapter is devoted to the description, analysis, approximations, and applications of the classical Mumford and Shah functional proposed for image segmentation. In [61–63], David Mumford and Jayant Shah have formulated an energy minimization problem that allows to compute optimal piecewise-smooth or piecewise-constant approximations u of a given initial image g. Since then, their model has been analyzed and considered in depth by many authors by studying properties of minimizers, approximations, and applications to image segmentation, image partition, image restoration, and more generally image analysis and computer vision.

Let $\Omega \subset \mathbb{R}^d$ be the image domain (an interval if $d = 1$; a rectangle in the plane if $d = 2$; or a rectangular parallelepiped if $d = 3$). More generally, it is assumed that Ω is open, bounded, and connected. Let $g{:}\Omega \to \mathbb{R}$ be a given grayscale image (a signal in one dimension, a planar image in two dimensions, or a volumetric image in three dimensions). It is natural and without losing any generality to assume that g is a bounded function in Ω, thus $g \in L^\infty(\Omega)$.

As formulated by Mumford and Shah [63], the *segmentation problem* in image analysis and computer vision consists in computing a decomposition

$$\Omega = \Omega_1 \cup \Omega_2 \cup \ldots \cup \Omega_n \cup K$$

of the domain of the image g such that

1. The image g varies smoothly and/or slowly *within* each Ω_i.
2. The image g varies discontinuously and/or rapidly across most of the boundary K between different Ω_i.

From the point of view of approximation theory, the segmentation problem may be restated as seeking ways to define and compute *optimal approximations* of a general function $g(x)$ by piecewise-smooth functions $u(x)$, i.e., functions u whose restrictions u_i to the pieces Ω_i of a decomposition of the domain Ω are continuous or differentiable.

In what follows, Ω_i will be disjoint connected open subsets of a domain Ω, each one with a piecewise-smooth boundary, and K will be a closed set, as the union of boundaries of Ω_i inside Ω, thus

$$\Omega = \Omega_1 \cup \Omega_2 \cup \ldots \cup \Omega_n \cup K, \quad K = \Omega \cap (\partial\Omega_1 \cup \ldots \cup \partial\Omega_n).$$

The functional E to be minimized for image segmentation is defined by [61–63],

$$E(u, K) = \mu^2 \int_\Omega (u - g)^2 dx + \int_{\Omega/K} |\nabla u|^2 \, dx + \nu \, |K|, \tag{1}$$

where $u : \Omega \to \mathbb{R}$ is continuous or even differentiable inside each Ω_i (or $u \in H^1(\Omega_i)$) and may be discontinuous across K. Here, $|K|$ stands for the total surface measure of the hypersurface K (the counting measure if $d = 1$, the length measure if $d = 2$, the area measure if $d = 3$). Later, $|K|$ will be defined by $\mathcal{H}^{d-1}(K)$, the $d - 1$ dimensional Hausdorff measure in $\mathbb{R}^d$.

As explained by Mumford and Shah, dropping any of these three terms in (1), inf $E = 0$: without the first, take $u = 0$, $K = 0$; without the second, take $u = g$, $K = 0$; without the third, take, for example, in the discrete case K to be the boundary of all pixels of the image g, each Ω_i be a pixel and u to be the average (value) of g over each pixel. The presence of all three terms leads to nontrivial solutions u, and an optimal pair (u, K) can be seen as a cartoon of the actual image g, providing a simplification of g.

An important particular case is obtained when E is restricted to piecewise-constant functions u, i.e., u = constant c_i on each open set Ω_i. Multiplying E by μ^{-2} gives

$$\mu^{-2} E(u, K) = \sum\nolimits_i \int_{\Omega_i} (g - c_i)^2 dx + \nu_0 \, |K|,$$

where $\nu_0 = \nu/\mu^2$. It is easy to verify that this is minimized in the variables c_i by setting

$$c_i = \text{mean}_{\Omega_i}(g) = \frac{\int_{\Omega_i} g(x) dx}{|\Omega_i|},$$

where $|\Omega_i|$ denotes here the Lebesgue measure of Ω_i (e.g., area if $d = 2$, volume if $d = 3$), so it is sufficient to minimize

$$E_0(K) = \sum\nolimits_i \int_{\Omega_i} (g - \text{mean}_{\Omega_i} g)^2 dx + v_0 |K| .$$

It is possible to interpret E_0 as the limit functional of E as $\mu \to 0$ [63].

Finally, the Mumford and Shah model can also be seen as a deterministic refinement of Geman and Geman's image restoration model [41].

2 Background: The First Variation

In order to better understand, analyze, and use the minimization problem (1), it is useful to compute its first variation with respect to each of the unknowns.

The definition of Sobolev functions $u \in W^{1,2}(U)$ [1] is now recalled, necessary to properly define a minimizer u when K is fixed.

Definition 1. Let $U \subset \mathbb{R}^d$ be an open set. Let $W^{1,2}(U)$ (or $H^1(U)$) denote the set of functions $u \in L^2(\Omega)$, whose first-order distributional partial derivatives belong to $L^2(U)$. This means that there are functions $u_1, \ldots, u_d \in L^2(U)$ such that

$$\int_U u(x) \frac{\partial \phi}{\partial x_i}(x) dx = - \int_U u_i(x) \phi(x) dx$$

for $1 \leq i \leq d$ and for all functions $\phi \in C_c^\infty(U)$.

Let $\frac{\partial u}{\partial x_i}$ denote the distributional derivative u_i of u and $\nabla u = \left(\frac{\partial u}{\partial x_1}, \ldots, \frac{\partial u}{\partial x_d} \right)$ its distributional gradient. In what follows, $|\nabla u|(x)$ denotes the Euclidean norm of the gradient vector at x. $H^1(U) = W^{1,2}(U)$ becomes a Banach space endowed with the norm

$$\|u\|_{W^{1,2}(U)} = \left[\int_U u^2 dx + \sum\nolimits_{i=1}^d \int_U \left(\frac{\partial u}{\partial x_i} \right)^2 dx \right]^{1/2} .$$

Minimizing in *u* with *K* Fixed

K is assumed to be fixed, as a closed subset of the open and bounded set $\Omega \subset \mathbb{R}^d$, and denote by

$$E(u) = \mu^2 \int_{\Omega/K} (u - g)^2 dx + \int_{\Omega/K} |\nabla u|^2 \, dx,$$

for $u \in W^{1,2}(\Omega\backslash K)$, where $\Omega\backslash K$ is open and bounded, and $g \in L^2(\Omega\backslash K)$. The following classical results are obtained as consequences of the standard method of calculus of variations.

Proposition 1. *There is a unique minimizer of the problem*

$$\inf_{u\in W^{1,2}(\Omega/K)} E(u). \tag{2}$$

Proof ([38]). First, note that $0 \le \inf E < +\infty$, since choosing $u_0 \equiv 0$, then $E(u_0) = \mu^2 \int_{\Omega\backslash K} g^2(x)dx < +\infty$. Thus, let $m = \inf_u E(u)$ and $\{u_j\}_{j\ge 1} \in W^{1,2}(\Omega\backslash K)$ be a minimizing sequence such that $\lim_{j\to\infty} E(u_j) = m$.

Recall that for $u, v \in L^2$,

$$\left\|\frac{u+v}{2}\right\|_2^2 + \left\|\frac{u-v}{2}\right\|_2^2 = \frac{1}{2}\|u\|_2^2 + \frac{1}{2}\|v\|_2^2,$$

and so

$$\left\|\frac{u+v}{2}\right\|_2^2 = \frac{1}{2}\|u\|_2^2 + \frac{1}{2}\|v\|_2^2 - \left\|\frac{u-v}{2}\right\|_2^2. \tag{3}$$

Let $u, v \in W^{1,2}(\Omega\backslash K)$, thus $E(u), E(v) < \infty$, and apply (3) to $u - g$ and $v - g$, and then to ∇u and ∇v; then

$$\begin{aligned} E\left(\tfrac{u+v}{2}\right) &= \tfrac{1}{2}E(u) + \tfrac{1}{2}E(v) - \tfrac{\mu^2}{4}\int_{\Omega/K}|u-v|^2\,dx - \tfrac{1}{4}\int_{\Omega/K}|\nabla(u-v)|^2\,dx \\ &= \tfrac{1}{2}E(u) + \tfrac{1}{2}E(v) - \begin{cases} \frac{\mu^2}{4}\|u-v\|^2_{W^{1,2}(\Omega/K)} + \left(1-\frac{\mu^2}{4}\right)\|\nabla(u-v)\|_2^2 \text{ if} \frac{1}{4} \ge \frac{\mu^2}{4} \\ \frac{1}{4}\|u-v\|^2_{W^{1,2}(\Omega/K)} + \left(\frac{\mu^2}{4}-1\right)\|u-v\|_2^2 \text{ if} \frac{1}{4} \le \frac{\mu^2}{4} \end{cases}. \end{aligned} \tag{4}$$

Choosing $u, v \in W^{1,2}(\Omega\backslash K)$, such that $E(u), E(v) \le m + \varepsilon$, then

$$\begin{aligned} m \le E\left(\tfrac{u+v}{2}\right) \le m + \varepsilon - \\ \begin{cases} \frac{\mu^2}{4}\|u-v\|^2_{W^{1,2}(\Omega/K)} + \left(1-\frac{\mu^2}{4}\right)\|\nabla(u-v)\|_2^2 \text{ if} \frac{1}{4} \ge \frac{\mu^2}{4} \\ \frac{1}{4}\|u-v\|^2_{W^{1,2}(\Omega/K)} + \left(\frac{\mu^2}{4}-1\right)\|u-v\|_2^2 \text{ if} \frac{1}{4} \le \frac{\mu^2}{4} \end{cases} \end{aligned}$$

thus,

$$\|u-v\|^2_{W^{1,2}(\Omega/K)} \le \begin{cases} \frac{4\varepsilon}{\mu^2} \text{if} \frac{1}{4} \ge \frac{\mu^2}{4} \\ 4\varepsilon \text{ if} \frac{1}{4} \le \frac{\mu^2}{4} \end{cases}. \tag{5}$$

Let $w_j = u_j - u_1$. From (5), $\{w_j\}$ is a Cauchy sequence in $W^{1,2}(\Omega\backslash K)$; let w denote its limit and set $u_0 = u_1 + w$. Then,

$$\begin{aligned} E(u_0) = \mu^2 \|u_0 - g\|_2^2 + \|\nabla u_0\|_2^2 &= \mu^2 \|(u_1 - g) + w\|_2^2 + \|\nabla u_1 + \nabla w\|_2^2 \\ &= \lim_{j\to+\infty} \left[\mu^2 \left\|(u_1 - g) + w_j\right\|_2^2 + \left\|\nabla u_1 + \nabla w_j\right\|_2^2\right] \\ &= \lim_{j\to+\infty} E(u_j) = m, \end{aligned}$$

by the continuity of L^2 norms. This shows the existence of minimizers. The uniqueness follows from (5) by taking $\varepsilon = 0$.

Proposition 2. *The unique solution u of (2) is solution of the elliptic problem*

$$\int_{\Omega/K} \nabla u(x) \cdot \nabla v(x) dx = -\mu^2 \int_{\Omega/K} [u(x) - g(x)]\, v(x) dx, \quad \forall v \in W^{1,2}(\Omega/K), \tag{6}$$

or of

$$\Delta u = \mu^2 (u - g)$$

in the sense of distributions in $\Omega\backslash K$, with associated boundary condition $\frac{\partial u}{\partial \vec{N}} = 0$ on $\partial(\Omega\backslash K)$, where $\vec{N}$ is the exterior unit normal to the boundary.

Proof. Indeed, let $\varepsilon \mapsto A(\varepsilon) = E(u + \varepsilon v)$ for $\int \in \mathbb{R}$ and arbitrary $v \in W^{1,2}(\Omega\backslash K)$. Then, A is a quadratic function of ε, given by

$$\begin{aligned} A(\varepsilon) = \mu^2 \textstyle\int_{\Omega/K} (u - g)^2 dx + \varepsilon^2 \mu^2 \int_{\Omega/K} v^2 dx + 2\varepsilon\mu^2 \int_{\Omega/K} (u - g) v dx \\ + \textstyle\int_{\Omega/K} |\nabla u|^2 dx + \varepsilon^2 \int_{\Omega/K} |\nabla v|^2 dx + 2\varepsilon \int_{\Omega/K} \nabla u \cdot \nabla v dx, \end{aligned}$$

with

$$\begin{aligned} A'(\varepsilon) = 2\varepsilon\mu^2 \textstyle\int_{\Omega/K} v^2 dx + 2\mu^2 \int_{\Omega/K} (u - g) v dx + 2\varepsilon \int_{\Omega/K} |\nabla v|^2 dx \\ + 2 \textstyle\int_{\Omega/K} \nabla u \cdot \nabla v dx, \end{aligned}$$

and

$$A'(0) = 2\mu^2 \int_{\Omega/K} (u - g) v dx + 2 \int_{\Omega/K} \nabla u \cdot \nabla v dx,$$

Since $E(u) = A(0) \leq A(\varepsilon) = E(u + \varepsilon v)$ for all $\int \in \mathbb{R}$ and all $v \in W^{1,2}(\Omega\backslash K)$, it implies that $A\prime(0) = 0$ for all such v, which yields the weak formulation (6).

If in addition u would be a strong classical solution of the problem, or if it would belong to $W^{2,2}(\Omega\backslash K)$, then integrating by parts in the last relation gives

$$A'(0) = 2\mu^2 \int_{\Omega/K} (u-g)v dx + 2\int_{\Omega/K} (\nabla u)v dx + 2\int_{\partial(\Omega/K)} \nabla u \cdot \vec{N} v dx = 0$$

Taking now $v \in C_0^1(\Omega\backslash K) \subset W^{1,2}(\Omega\backslash K)$ gives

$$\Delta u = \mu^2(u-g) \text{ in } \Omega/K.$$

Using this and taking now $v \in C^1(\Omega\backslash K)$, the associated implicit boundary condition $\nabla u.\vec{N} = \frac{\partial u}{\partial \vec{N}} = 0$ on the boundary of $\Omega\backslash K$ (in other words, on the boundary of Ω and of each Ω_i) is obtained.

Assume now that $g \in L^\infty(\Omega\backslash K)$, which is not a restrictive assumption when g represents an image. It can be shown that the unique minimizer u of (2) satisfies $||u||_\infty \leq ||g||_\infty$ (as expected, due to the smoothing properties of the energy). To prove this, first the following classical lemma is stated (see, e.g., Ref. [38], Chapter A3).

Lemma 1. *If $\Omega\backslash K$ is open, and if $u \in W^{1,2}(\Omega\backslash K)$, then $u^+ = \max(u, 0)$ also lies in $W^{1,2}(\Omega\backslash K)$ and $|\nabla u^+(x)| \leq |\nabla u(x)|$ almost everywhere.*

Now let $u^*(x) = max\{-||g||_\infty, \min(||g||_\infty, u(x))\}$ be the obvious truncation of u. Lemma 1 implies that $u^* \in W^{1,2}(\Omega\backslash K)$ and that $\int_{\Omega\backslash K} |\nabla u^*(x)|^2 dx \leq \int_{\Omega\backslash K} |\nabla u(x)|^2 dx$. Obviously, also $\int_{\Omega\backslash K} (u^* - g)^2 dx \leq \int_{\Omega\backslash K} (u-g)^2 dx$, thus $E(u^*) \leq E(u)$. But u is the unique minimizer of E; therefore, $u(x) = u^*(x)$ almost everywhere and $||u||_\infty \leq ||g||_\infty$.

Remark 1. Several classical regularity results for a weak solution u of (2) can be stated:

- If $g \in L^\infty(\Omega\backslash K)$, then $u \in C^1_{\text{loc}}(\Omega\backslash K)$ (see e.g., Ref. [38], Chapter A3).
- If $g \in L^2(\Omega\backslash K)$, then $u \in W^{2,2}_{\text{loc}}(\Omega\backslash K) = H^2_{\text{loc}}(\Omega\backslash K)$, which implies that u solves the PDE (see, e.g., Ref. [39], Chapter 6.3).

$$\Delta u = \mu^2(u-g) \text{ a.e. in } \Omega/K.$$

3 Minimizing in *K*

Here we formally compute the first variation of $E(u, K)$ with respect to K. Let us assume that (u, K) is a minimizer of E from (1), and we vary K. Let us assume that locally, $K \cap U$ is the graph of a regular function ϕ, where U is a small neighborhood near a regular, simple point P of K.

Without loss of generality, it can be assumed that $U = D \times I$ where I is an interval in $\mathbb{R}$ and $K \cap U = \{(x_1, x_2, \ldots, x_d) \in U = D \times I : x_d = \phi(x_1, \ldots, x_{d-1})\}$. Let u^+ denote the restriction of u to

$$U^+ = \{(x_1, x_2, \ldots, x_d) : x_d > \phi(x_1, \ldots, x_{d-1})\} \cap U,$$

and u^- the restriction of u to

$$U^- = \{(x_1, x_2, \ldots, x_d) : x_d > \phi(x_1, \ldots, x_{d-1})\} \cap U,$$

and choose H^1 extensions of u^+ from U^+ to U, and of u^- from U^- to U. For small ε, define a deformation K_ε of K inside U as the graph of

$$x_d = \phi(x_1, \ldots, x_{d-1}) + \varepsilon\psi(x_1, \ldots, x_{d-1}),$$

such that ψ is regular and zero outside D, and $K_\varepsilon = K$ outside U. Define

$$u_\varepsilon(x) = \begin{cases} u(x) & if \quad x \notin U, \\ (\text{extension of } u^+)(x) & if \quad x \in U, x \text{ above } K_\varepsilon \cap U \\ (\text{extension of } u^-)(x) & if \quad x \in U, x \text{ above } K_\varepsilon \cap U. \end{cases}$$

Now, using $z = (x_1, \ldots, x_{d-1})$,

$$\begin{aligned} E(u_\varepsilon, K_\varepsilon) - E(u, K) &= \mu^2 \int_U \left[(u_\varepsilon - g)^2 dx - (u - g)^2\right] dx \\ &\quad + \int_{U/K_\varepsilon} |\nabla u_\varepsilon|^2 \, dx - \int_{U/K} |\nabla u|^2 \, dx + \nu \left[|K_\varepsilon \cap U| - |K \cap U|\right] \\ &= \mu^2 \int_D \left(\int_{\phi(z)}^{\phi(z)+\varepsilon\psi(z)} \left[(u^- - g)^2 - (u^+ - g)^2\right] dx_d \right) dz \\ &\quad + \int_D \left(\int_{\phi(z)}^{\phi(z)+\varepsilon\psi(z)} \left[|\nabla u^-|^2 - \left|\nabla u^+\right|^2\right] dx_d \right) dz \\ &\quad + \nu \int_D \left[\sqrt{1 + |\nabla(\phi + \varepsilon\psi)|^2} - \sqrt{1 + |\nabla\phi|^2} \right] dz \end{aligned}$$

Thus,

$$\begin{aligned} \lim_{\varepsilon \to 0} \tfrac{E(u_\varepsilon, K_\varepsilon) - E(u, K)}{\varepsilon} &= \mu^2 \int_D \left[(u^- - g)^2 - (u^+ - g)^2\right] \big|_{x_d = \phi(z)} \psi(z) dz \\ + \int_D \left[|\nabla u^-|^2 - \left|\nabla u^+\right|^2\right] \Big|_{x_d = \phi(z)} \psi(z) dz &+ \nu \int_D \tfrac{\nabla\phi \cdot \nabla\psi}{\sqrt{1 + |\nabla\phi|^2}} dz = \end{aligned}$$

for all such ψ, since (u, K) is a minimizer. Integrating by parts, we formally obtain for all ψ:

$$\int_D \left\{ \left[\left(\mu^2 (u^- - g)^2 + |\nabla u^-|^2\right) - \left(\mu^2 (u^+ - g)^2 + \left|\nabla u^+\right|^2\right) \right] \big|_{x_d = \phi(z)} - \nu \operatorname{div} \left(\tfrac{\nabla\phi}{\sqrt{1 + |\nabla\phi|^2}} \right) \right\} \psi(z) dz = 0,$$

and obtain the first variation with respect to K,

$$\left[\mu^2(u^- - g)^2 + |\nabla u^-|^2\right] - \left[\mu^2(u^+ - g)^2 + \left|\nabla u^+\right|^2\right] - \nu \operatorname{div}\left(\frac{\nabla \phi}{\sqrt{1 + |\nabla \phi|^2}}\right) = 0 \tag{7}$$

on $K \cap U$. Noticing that the last term represents the curvature of $K \cap U$, and if we write the energy density as

$$e(u; x) = \mu^2(u(x) - g(x))^2 + |\nabla u(x)|^2,$$

we finally obtain

$$e(u^+) - e(u^-) + \nu \operatorname{curv}(K) = 0 \text{ on } K$$

(at regular points of K, provided that the traces of u and of $|\nabla u|$ on each side of K are taken in the sense of Sobolev traces).

This section is concluded by stating another important result from [63] regarding the type of singular points of K, when (u, K) is a minimizer of E from (1), in two dimensions, $d = 2$. For the rather technical proof of this result, the reader is referred to the instructive and inspiring constructions from [63].

Theorem 1. *Let $d = 2$. If (u, K) is a minimizer of $E(u, K)$ such that K is a union of simple $C^{1,1}$-curves K_i meeting $\partial\Omega$ and meeting each other only at their endpoints, then the only vertices of K are:*

1. *Points P on the boundary $\partial\Omega$ where one K_i meets $\partial\Omega$ perpendicularly*
2. *Triple points P where three K_i meet with angles $2\pi/3$*
3. *Crack tips where a K_i ends and meets nothing.*

The later sections will discuss cases when the minimizer u is restricted to a specific class of piecewise-constant or piecewise-smooth functions.

4 Mathematical Modeling and Analysis: The Weak Formulation of the Mumford and Shah Functional

To better study the mathematical properties of the Mumford and Shah functional (1), it is necessary to define the measure of K as its $d-1$-dimensional Hausdorff measure $\mathcal{H}^{d-1}(K)$, which is the most natural way to extend the notion of length to nonsmooth sets. Recall the definition of the Hausdorff measure [4, 38, 40].

Definition 2. For $K \subset \mathbb{R}^d$ and $n > 0$, set

$$\mathcal{H}^n(K) = \sup_{\varepsilon > 0} \mathcal{H}^n_\varepsilon(K),$$

called the n-dimensional Hausdorff measure of the set K, where

$$\mathcal{H}^n_\varepsilon(K) = c_n \inf\left\{\sum\nolimits_{i=1}^{\infty} (\text{diam } A_i)^n\right\},$$

where the infimum is taken over all countable families $\{A_i\}_{i=1}^{\infty}$ of open sets A_i such that

$$K \subset \cup^{i=1}{}^{\infty} A_i \text{ and diam } A_i \leq \varepsilon \text{ for all } i.$$

Here, the constant c_n is chosen so that $\mathcal{H}^n$ coincides with the Lebesgue measure on n planes.

Remark 2. When n is an integer and K is contained in a C^1-surface of dimension n, $\mathcal{H}^n(K)$ coincides with its n-dimensional surface measure.

Consider a first variant of the functional,

$$E(u, K) = \mu^2 \int_{\Omega/K} (u-g)^2 dx + \int_{\Omega/K} |\nabla u|^2 \, dx + \nu \mathcal{H}^{d-1}(K). \tag{8}$$

In order to apply the direct method of calculus of variations for proving existence of minimizers, it is necessary to find a topology for which the functional is lower semicontinuous, while ensuring compactness of minimizing sequences. Unfortunately, the last functional $K \mapsto \mathcal{H}^{d-1}(K)$ is not lower semicontinuous with respect to any compact topology [4, 8, 38]. To overcome this difficulty, the set K is substituted by the jump set S_u of u, thus K is eliminated, and the problem, called the weak formulation, becomes, in its second variant,

$$\inf_u \left\{ F(u) = \mu^2 \int_{\Omega/S_u} (u-g)^2 dx + \int_{\Omega/S_u} |\nabla u|^2 \, dx + \nu \mathcal{H}^{d-1}(S_u) \right\}. \tag{9}$$

For illustration, also given is the weak formulation in one dimension, for signals. The problem of reconstructing and segmenting a signal u from a degraded input g deriving from a distorted transmission can be modeled as finding the minimum

$$\inf_u \left\{ \mu^2 \int_a^b (u-g)^2 dt + \int_{(a,b)/S_u} \left|u'\right|^2 dt + \nu \#(S_u) \right\}.$$

where $\Omega = (a, b)$, S_u denotes the set of discontinuity points of u in the interval (a, b), and $\#(S_u) = \mathcal{H}^0(S_u)$ denotes the counting measure of S_u or its cardinal.

In order to show that (9) has a solution, the following notion of special functions of bounded variation and the following important lemma due to Ambrosio [3,4] are necessary.

Definition 3. A function $u \in L^1(\Omega)$ is a special function of bounded variation on Ω if its distributional derivative can be written as

$$Du = \nabla u dx + (u^+ - u^-)\vec{N}_u \, \mathcal{H}^{d-1}\big|_{S_u}$$

such that $\nabla u \in L^1(\Omega)$, S_u is of finite Hausdorff measure, $(u^+ - u^-)\vec{N}_{u\chi S_u} \in L^1\left(\Omega, \mathcal{H}^{d-1}|_{S_u}, \mathbb{R}^d\right)$, where u^+ and u^- are the traces of u on each side of the jump part S_u, and $\vec{N}_u$ is the unit normal to S_u. The space of special functions of bounded variation is denoted by SBV(Ω).

Lemma 2. *Let $u_n \in$ SBV(Ω) be a sequence of functions such that there exists a constant $C > 0$ with $|u_n(x)| \leq C < \infty$ a.e. $x \in \Omega$ and $\int_\Omega |\nabla u_n|^2 \, dx + \mathcal{H}^{d-1}(S_{u_n}) \leq C$. Then, there exists a subsequence u_{nk} converging a.e. to a function $u \in$ SBV(Ω). Moreover, ∇u_{nk} converges weakly in $L^2(\Omega)^d$ to ∇u, and*

$$\mathcal{H}^{d-1}(S_u) \leq \liminf_{n_k \to \infty} \mathcal{H}^{d-1}(S_{u_{n_k}}).$$

Theorem 2. *Let $g \in L^\infty(\Omega)$, with $\Omega \subset \mathbb{R}^d$ open, bounded, and connected. There is a minimizer $u \in$ SBV(Ω) $\cap L^\infty(\Omega)$ of*

$$F(u) = \mu^2 \int_{\Omega/S_u} (u-g)^2 dx + \int_{\Omega/S_u} |\nabla u|^2 \, dx + \nu \mathcal{H}^{d-1}(S_u).$$

Proof. Notice that $0 \leq \inf^{\infty}_{\mathrm{SBV}(\Omega)\cap L(\Omega)} F < \infty$, because we can take $u_0 = 0 \in$ SBV(Ω) $\cap L^\infty(\Omega)$ and using the fact that $g \in L^\infty(\Omega) \subset L^2(\Omega)$, $F(u_0) < \infty$. Thus, there is a minimizing sequence $u_n \in$ SBV(Ω) $\cap L^\infty(\Omega)$ satisfying $\lim_{n\to\infty} F(u_n) = \inf F$. Also notice that, by the truncation argument from before, it can be assumed that $||u_n||_\infty \leq ||g||_\infty < \infty$. Since $F(u_n) \leq C < \infty$ for all $n \geq 0$, and using $g \in L^\infty(\Omega) \subset L^2(\Omega)$, we deduce that $||u_n||_2 \leq C$ and $\int_{\Omega/S_{u_n}} |\nabla u_n|^2 \, dx + \mathcal{H}^{d-1}(S_{u_n}) < C$ for some positive real constant C. Using these and Ambrosio's compactness result, it can be deduced that there is a subsequence u_{nk} of u_n, and $u \in$ SBV(Ω), such that $u_{n_k} \rightharpoonup u$ in $L^2(\Omega)$, $\nabla u_{n_k} \rightharpoonup \nabla u$ in $L^2(\Omega)^d$. Therefore, $F(u) \leq \liminf_{nk\to\infty} F(u_{nk}) = \inf F$, and it can also be deduced that $||u||_\infty \leq ||g||_\infty$.

For additional existence, regularity results and fine properties of minimizers, and for the connections between problems (8) and (9), the reader is referred to Dal Maso et al. [55, 56], the important monographs by Morel and Solimini [60], by Chambolle [26], by Ambrosio et al. [4], by David [38], and by Braides [19]. Existence and regularity of minimizers for the piecewise-constant case can be found in [63], Congedo and Tamanini [52, 57, 75, 76], and Larsen [51], among other works.

5 Numerical Methods: Approximations to the Mumford and Shah Functional

Since the original Mumford and Shah functional (1) (or its weak formulation (9)) is nonconvex, it has an unknown set K of lower dimension and is not the lower-semicontinuous envelope of a more amenable functional, and it is difficult to find smooth approximations and to solve the minimization in practice. Several approximations have been proposed, including the weak membrane model and the graduate non-convexity of Blake and Zisserman [16] (which can be seen as a discrete version of the Mumford and Shah segmentation problem); discrete finite differences approximations starting with the work of Chambolle [23–25] (also proving the Γ-convergence of Blake-Zisserman approximations to the weak Mumford-Shah functional in one dimension); finite element approximations by Chambolle and Dal Maso [27] and by Chambolle and Bourdin [17, 18]; phase-field elliptic approximations due to Ambrosio and Tortorelli [5, 6] (with generalizations presented by [19] and extensions by Shah [74], and Alicandro et al. [2]); region growing and merging methods proposed by Koepfler et al. [48], by Morel and Solimini [60], and by Dal Maso et al. [55,56] and level set approximations proposed by Chan and Vese [29–32, 79], by Samson et al. [71], and by Tsai et al. [78]; and approximations by nonlocal functionals by Braides and Dal Maso [20], among other approximations. Presented in this section in much more detail are the phase-field elliptic approximations and the level set approximations together with their applications.

For proving the convergence of some of these approximations to the Mumford and Shah functional, the notion of Γ-convergence is used, which is briefly recalled below. The interested reader is referred to Dal Maso [54] for a comprehensive introduction to Γ-convergence.

The reader is referred to the monographs and textbooks by Braides [19], by Morel and Solimini [60], and by Ambrosio et al. [4] on detailed presentations of approximations to the Mumford and Shah functional.

Definition 4. Let $X = (X, D)$ be a metric space. We say that a sequence $F_j : X \to [-\infty, +\infty]$ Γ-converges to $F : X \to [-\infty, +\infty]$ (as $j \to \infty$) if for all $u \in X$ we have

1. (lim inf inequality) for every sequence $(u_j) \subset X$ converging to u,

$$F(u) \leq \liminf_j F_j(u_j) \tag{10}$$

2. (existence of a recovery sequence) there exists a sequence $(u_j) \subset X$ converging to u such that

$$F(u) \geq \limsup_j F_j(u_j),$$

or, equivalently by (10),

$$F(u) = \lim_j F_j(u_j).$$

The function F is called the Γ-limit of (F_j) (with respect to D), and we write $F = \Gamma\text{-}\lim_j F_j$.

The following fundamental theorem is essential in the convergence of some of the approximations.

Theorem 3 (Fundamental Theorem of Γ-convergence). *Let us suppose that $F = \Gamma\text{-}\lim_j F_j$, and let a compact set $C \subset X$ exist such that $\inf_X F_j = \inf_C F_j$ for all j. Then, there is minimum of F over X such that*

$$\min_X F = \lim_j \inf_X F_j,$$

and if $(u_j) \subset X$ is a converging sequence such that $\lim_j F_j\ (u_j) = \lim_j \inf_X F_j$, then its limit is a minimum point of F.

6 Ambrosio and Tortorelli Phase-Field Elliptic Approximations

A specific strategy, closer to the initial formulation of the Mumford-Shah problem in terms of pairs $(u, K = S_u)$, is based on the approximation by functionals depending on two variables (u, v), the second one related to the set $K = S_u$.

7 Approximations of the Perimeter by Elliptic Functionals

The Modica-Mortola theorem [58, 59] enables the variational approximation of the perimeter functional $E \mapsto P(E, \Omega) = \int_\Omega |D_{\chi E}| < \infty$ of an open subset E of Ω by the quadratic, elliptic functionals

$$MM_\varepsilon(v) = \int_\Omega \left(\varepsilon |\nabla v|^2 + \frac{W(v)}{\varepsilon} \right) dx, \quad v \in W^{1,2}(\Omega),$$

where $W(t)$ is a "double-well" potential. For instance, choosing $W(t) = t^2(1-t)^2$, assuming that Ω is bounded with Lipschitz boundary and setting $MM_\varepsilon(v) = \infty$ if $v \in L^2(\Omega) \setminus W^{1,2}(\Omega)$, the functionals $MM_\varepsilon(v)$ Γ-converge in $L^2(\Omega)$ to

$$F(v) = \begin{cases} \frac{1}{3} P(E, \Omega) & if\ v = \chi_E \text{ for some } E \in \mathcal{B}(\Omega), \\ \infty & \text{otherwise,} \end{cases}$$

where $\mathcal{B}(\Omega)$ denotes the σ-algebra of Borel subsets of Ω.

Minimizing the functional $MM_\varepsilon(v)$ with respect to v yields the associated Euler-Lagrange equation and boundary condition,

$$W'(v) = 2\varepsilon^2 \Delta v \text{ in } \Omega, \qquad \frac{\partial v}{\partial \vec{N}} = 0 \text{ on } \partial\Omega,$$

which can be easily solved in practice by finite differences.

8 Ambrosio-Tortorelli Approximations

In the Mumford and Shah functional, the set $K = S_u$ is not necessarily the boundary of an open and bounded domain, but a construction similar to $MM_\varepsilon(v)$ can still be used, with the potential $W(t) = \frac{1}{4}(1-t)^2$ instead. Ambrosio and Tortorelli proposed two elliptic approximations [5,6] to the weak formulation of the Mumford and Shah problem. Presented is the second one [6], being simpler than the first one [5], and commonly used in practice.

Let $X = L^2(\Omega)^2$ and let us define

$$AT_\varepsilon(u,v) = \int_\Omega (u-g)^2 dx + \beta \int_\Omega v^2 |\nabla u|^2 dx + \alpha \int_\Omega \left(\varepsilon |\nabla v|^2 + \frac{(v-1)}{4\varepsilon} \right) dx \tag{11}$$

if $(u, v) \in W^{1,2}(\Omega)^2$, $0 \le v \le 1$, and $AT_\varepsilon(u, v) = +\infty$ otherwise.

Also defined is the limiting Mumford-Shah functional,

$$F(u,v) = \begin{cases} \int_\Omega (u-g)^2 dx + \beta \int_\Omega |\nabla u|^2 + \alpha \mathcal{H}^{d-1}(S_u) & \text{if } u \in \mathrm{SBV}(\Omega), \quad v \equiv 1, \\ +\infty & \text{otherwise.} \end{cases}$$

Theorem 4. *AT_ε Γ-converges to F as $\varepsilon \to 0$ in $L^2(\Omega)$. Moreover, AT_ε admits a minimizer $(u_\varepsilon, v_\varepsilon)$ such that up to subsequences, u_ε converges to some $u \in \mathrm{SBV}(\Omega)$ a minimizer of $F(u, 1)$ and* $\inf AT_\varepsilon(u_\varepsilon, v_\varepsilon) \to F(u, 1)$.

Interesting generalizations of this result are given and proved by Braides in [19].

In practice, the Euler-Lagrange equations associated with the alternating minimization of AT_ε with respect to $u = u_\varepsilon$ and $v = v_\varepsilon$ are used and discretized by finite differences. These are

$$\frac{\partial AT_\varepsilon(u,v)}{\partial u} = 2(u-g) - 2\beta \operatorname{div}(v^2 \nabla u) = 0$$
$$\frac{\partial AT_\varepsilon(u,v)}{\partial v} = 2\beta v |\nabla u|^2 - 2\alpha\varepsilon \Delta v + \frac{\alpha}{2\varepsilon}(v-1).$$

One of the finite difference approximations to compute u and v in two dimensions $x = (x_1, x_2)$ is as follows. A time-dependent scheme is used in $u = u(x_1, x_2, t)$ and

a stationary semi-implicit fixed-point scheme in $v = v(x_1, x_2)$. Let $\Delta x_1 = \Delta x_2 = h$ be the step space, Δt be the time step, and $g_{i,j}$, $u^n_{i,j}$, $v^n_{i,j}$ be the discrete versions of g, and of u and v at iteration $n \geq 0$, for $1 \leq i \leq M$, $1 \leq j \leq N$. Initialize $u^0 = g$ and $v^0 = 0$.

For $n \geq 1$, compute and repeat to steady state, for $i = 2, \ldots, M-1$ and $j = 2, \ldots, N-1$ (combined with Neumann boundary conditions on $\partial\Omega$):

$$
\begin{aligned}
|\nabla u^n|^2_{i,j} &= \left(\tfrac{u^n_{i+1,j}-u^n_{i,j}}{h}\right)^2 + \left(\tfrac{u^n_{i,j+1}-u^n_{i,j}}{h}\right)^2,\\
0 &= 2\beta v^{n+1}_{i,j}\,|\nabla u^n|^2_{i,j} - 2\tfrac{\alpha\varepsilon}{h^2}\left(v^n_{i+1,j} + v^n_{i-1,j} + v^n_{i,j+1} + v^n_{i,j-1}\right.\\
&\quad \left. -4v^{n+1}_{i,j}\right) + \tfrac{\alpha}{2\varepsilon}\left(v^{n+1}_{i,j} - 1\right),\\
\tfrac{u^{n+1}_{i,j}-u^n_{i,j}}{\Delta t} &= -\left(u^n_{i,j} - g_{i,j}\right) + \tfrac{\beta}{h^2}\left[\left(v^{n+1}_{i,j}\right)^2\left(u^n_{i+1,j} - u^n_{i,j}\right) + \left(v^{n+1}_{i,j}\right)^2\right.\\
&\quad \left.\left(u^n_{i,j+1} - u^n_{i,j}\right) - \left(v^{n+1}_{i-1,j}\right)^2\left(u^n_{i,j} - u^n_{i-1,j}\right) - \left(v^{n+1}_{i,j-1}\right)^2\left(u^n_{i,j} - u^n_{i,j-1}\right)\right]
\end{aligned}
$$

which is equivalent with

$$
\begin{aligned}
|\nabla u^n|^2_{i,j} &= \left(\tfrac{u^n_{i+1,j}-u^n_{i,j}}{h}\right)^2 + \left(\tfrac{u^n_{i,j+1}-u^n_{i,j}}{h}\right)^2,\\
v^{n+1}_{i,j} &= \frac{\frac{\alpha}{2\varepsilon} + \frac{2\alpha\varepsilon}{h^2}\left(v^n_{i+1,j} + v^n_{i-1,j} + v^n_{i,j+1} + v^n_{i,j-1}\right)}{\frac{\alpha}{2\varepsilon} + 2\beta\,|\nabla u^n|^2_{i,j} + \frac{8\alpha\varepsilon}{h^2}},\\
u^{n+1}_{i,j} &= u^n_{i,j} + \Delta t\left\{-\left(u^n_{i,j} - g_{i,j}\right) + \tfrac{\beta}{h^2}\left[\left(v^{n+1}_{i,j}\right)^2\left(u^n_{i+1,j} - u^n_{i,j}\right)\right.\right.\\
&\quad + \left(v^{n+1}_{i,j}\right)^2\left(u^n_{i,j+1} - u^n_{i,j}\right) - \left(v^{n+1}_{i-1,j}\right)^2\left(u^n_{i,j} - u^n_{i-1,j}\right)\\
&\quad \left.\left. - \left(v^{n+1}_{i,j-1}\right)^2\left(u^n_{i,j} - u^n_{i,j-1}\right)\right]\right\}
\end{aligned}
$$

Presented are experimental results obtained using the above Ambrosio-Tortorelli approximations applied to the well-known Barbara image shown in Fig. 1 left. Segmented images u are shown in Fig. 2, and the corresponding edge sets v are shown in Fig. 3 for varying coefficients α, $\beta \in \{1, 5, 10\}$. We notice that less regularization (decreasing both α and β) gives more edges in v, as expected; thus, u is closer to g. Fixed α and increasing β give smoother image u and fewer edges in v. Keeping fixed β but varying α does not produce much variation in the results. Also shown in Fig. 1 right is the numerical energy versus iterations for the case $\alpha = \beta = 10$, $\varepsilon = 0.0001$.

Applications of the Ambrosio-Tortorelli approximations to image restoration will be presented in details in Sect. 5.

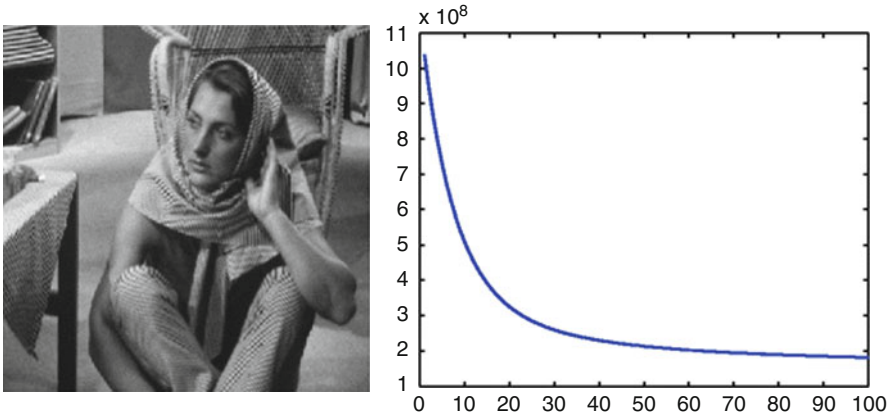


Fig. 1 *Left*: original image g. *Right*: numerical energy versus iterations for the Ambrosio-Tortorelli approximations ($\alpha = \beta = 10, \varepsilon = 0.0001$)

9 Level Set Formulations of the Mumford and Shah Functional

In this section are reviewed the level set formulations for minimizing the Mumford and Shah functional, as proposed initially by Chan and Vese [29–32,79], and by Tsai et al. [78] (see also the related work by Samson et al. [71] and Cohen et al. [36,37]). These make the link between curve evolution, active contours, and Mumford-Shah segmentation. These models have been proposed by restricting the set of minimizers u to specific classes of functions: piecewise constant, piecewise smooth, and with the edge set K represented by a union of curves or surfaces that are boundaries of open subsets of Ω. For example, if K is the boundary of an open-bounded subset of Ω, then it can be represented implicitly, as the zero-level line of a Lipschitz-continuous level set function. Thus, the set K as an unknown is substituted by an unknown function that defines it implicitly, and the Euler-Lagrange equations with respect to the unknowns can be easily computed and discretized.

Following the level set approach [67,68,72,73], let $\phi{:}\Omega \to \mathbb{R}$ be a Lipschitz-continuous function. Recalled is the variational level set terminology that will be useful to rewrite the Mumford and Shah functional in terms of (u, ϕ), instead of (u, K). This is inspired by the work of Zhao et al. [83] on a variational level set approach for the motion of triple junctions in the plane.

Used here is the one-dimensional (1D) Heaviside function H, defined by

$$H(z) = \begin{cases} 1 \text{ if } z \geq 0 \\ 0 \text{ if } z < 0 \end{cases},$$

and its distributional derivative $\delta = H'$ (in the weak sense). In practice, it may be necessary to work with smooth approximations of the Heaviside and δ functions.

Fig. 2 Piecewise-smooth images u as minimizers of the Ambrosio-Tortorelli approximations for $\varepsilon = 0.0001$ and various values of (α, β)

Used here are the following C^∞ approximations as $\varepsilon \to 0$ given by [29,31],

$$H_\varepsilon(z) = \frac{1}{2}\left[1 + \frac{2}{\pi}\arctan\left(\frac{z}{\varepsilon}\right)\right], \quad \delta_\varepsilon = H'_\varepsilon.$$

The area (or the volume) of the region $\{x \in \Omega\colon \phi(x) > 0\}$ is

$$A\{x \in \Omega : \phi(x) > 0\} = \int_\Omega H(\phi(x))dx,$$

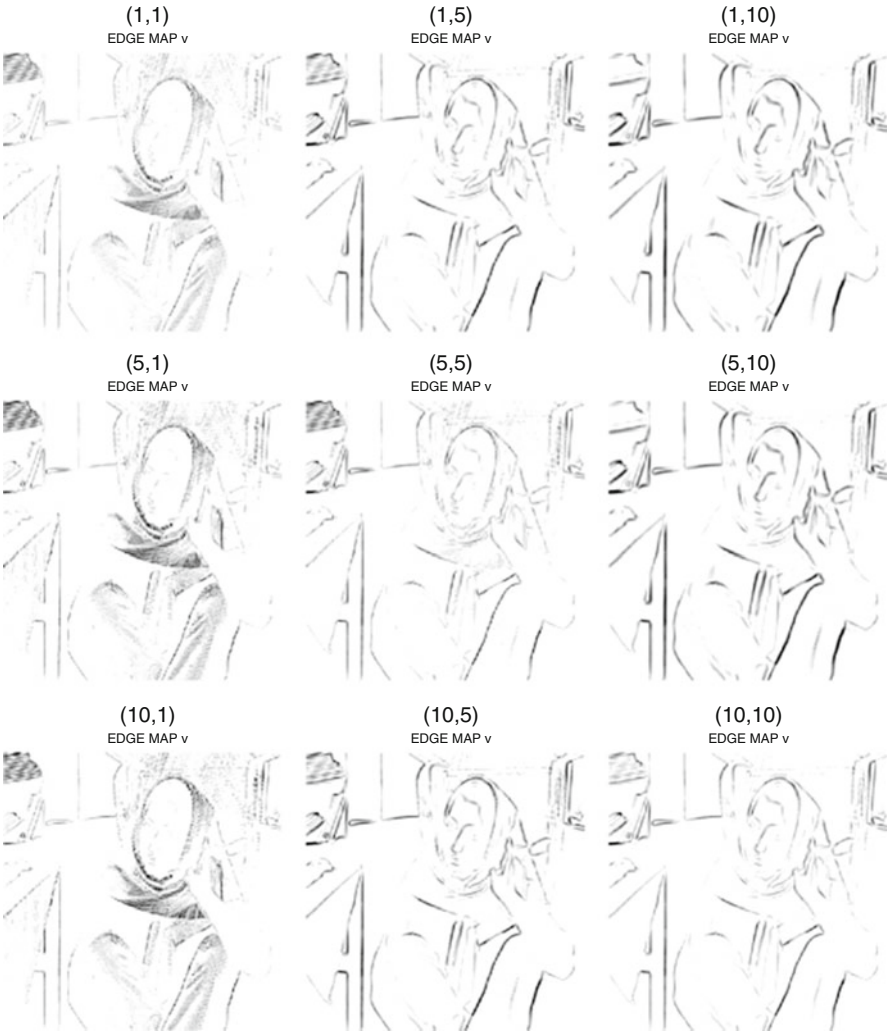


Fig. 3 Corresponding edge sets v as minimizers of the Ambrosio-Tortorelli approximations for $\varepsilon = 0.0001$ and various values of (α, β)

and for a level parameter $l \in \mathbb{R}$, the area (or volume) of the region $\{x \in \Omega\colon \phi(x) > l\}$ is

$$A\{x \in \Omega : \phi(x) > 0\} = \int_{\Omega} H(\phi(x) - l)dx.$$

The perimeter (or more generally the surface area) of the region $\{x \in \Omega\colon \phi(x) > 0\}$ is given by

$$L\{x \in \Omega : \phi(x) > 0\} = \int_{\Omega} |DH(\phi)|,$$

which is the total variation of $H(\phi)$ in Ω, and the perimeter (or surface area) of $\{x \in \Omega: \phi(x) > l\}$ is

$$L\{x \in \Omega : \phi(x) > l\} = \int_{\Omega} |DH(\phi - l)|.$$

Given the image data $g \in L^{\infty}(\Omega) \subset L^2(\Omega)$ to be segmented, the averages of g over the (nonempty) regions $\{x \in \Omega: \phi(x) > 0\}$ and $\{x \in \Omega: \phi(x) < 0\}$, respectively, are

$$\frac{\int_{\Omega} g(x)H(\phi(x))dx}{\int_{\Omega} H(\phi(x))\,dx} \text{ and } \frac{\int_{\Omega} g(x)(1 - H(\phi(x)))dx}{\int_{\Omega} (1 - H(\phi(x)))\,dx} = \frac{\int_{\Omega} g(x)H(-\phi(x))\,dx}{\int_{\Omega} H(-\phi(x))\,dx}.$$

More generally, for a given level parameter $l \in \mathbb{R}$, the averages of g over the corresponding (nonempty) regions $\{x \in \Omega: \phi(x) > l\}$ and $\{x \in \Omega: \phi(x) < l\}$, respectively, are

$$\frac{\int_{\Omega} g(x)H(\phi(x) - l)dx}{\int_{\Omega} H(\phi(x) - l)\,dx} \text{ and } \frac{\int_{\Omega} g(x)H(l - \phi(x))dx}{\int_{\Omega} H(l - \phi(x))\,dx}.$$

Proved next is that if H and δ are substituted by the above C^{∞} approximations H_{ε}, δ_{ε} as $\varepsilon \to 0$, approximations of the area and length (perimeter) measures are obtained. It is obviously found that $H_{\varepsilon}(z) \to H(z)$ for all $z \in \mathbb{R}$, as $\varepsilon \to 0$, and that the approximating area term $A_{\varepsilon}(\phi) = \int_{\Omega} H_{\varepsilon}(\phi(x))dx$ converges to $A(\phi) = \int_{\Omega} H(\phi(x))dx$.

Generalizing a result of Samson et al. [71], it can be shown [35] that our approximating functional $L_{\varepsilon}(\phi) = \int_{\Omega} |DH_{\varepsilon}(\phi)|dx = \int_{\Omega} \delta_{\varepsilon}(\phi)|\nabla\phi|dx$ converges to the length $|K|$ of the zero-level line $K = \{x \in \Omega: \phi(x) = 0\}$, under the assumption that $\phi:\Omega \to \mathbb{R}$ is Lipschitz. The same result holds for the case of any l-level curve of ϕ and not only for the 0-level curve.

Theorem 5. *Let us define*

$$L_{\varepsilon}(\phi) = \int_{\Omega} |\nabla H_{\varepsilon}(\phi)|\,dx = \int_{\Omega} \delta_{\varepsilon}(\phi)\,|\nabla\phi|\,dx.$$

Then, we have

$$\lim_{\varepsilon \to 0} L_{\varepsilon}(\phi) = \int_{\{\phi=0\}} ds = |K|,$$

where $K = \{x \in \Omega: \phi(x) = 0\}$.

Proof. Using co-area formula [40], the following is found:

$$L_\varepsilon(\phi) = \int_{\mathbb{R}} \left[\int_{\phi=\rho} \delta_\varepsilon(\phi(x)) ds \right] d\rho = \int_{\mathbb{R}} \left[\delta_\varepsilon(\rho) \int_{\phi=\rho} ds \right] d\rho.$$

By setting $h(\rho) = \int_{\phi=\rho} ds$, the following is obtained

$$L_\varepsilon(\phi) = \int_{\mathbb{R}} \delta_\varepsilon(\rho) h(\rho) d\rho = \int_{\mathbb{R}} \frac{1}{\pi} \frac{\varepsilon}{\varepsilon^2 + \rho^2} h(\rho) d\rho.$$

By the change of variable $\theta = \frac{\rho}{\varepsilon}$, the following is obtained

$$\begin{aligned} &\lim_{\varepsilon\to 0} L_\varepsilon(\phi) = \lim_{\varepsilon\to 0} \int_R \frac{1}{\pi} \frac{\varepsilon}{\varepsilon^2+\varepsilon^2\theta^2} h(\theta\varepsilon) d\theta = \lim_{\varepsilon\to 0} \int_{\mathbb{R}} \frac{1}{\pi} \frac{1}{1+\theta^2} h(\theta\varepsilon)\, d\theta \\ &= h(0) \int_{\mathbb{R}} \frac{1}{\pi} \frac{1}{1+\theta^2} d\theta = h(0) \frac{1}{\pi} \arctan\theta \Big|_{-\infty}^{+\infty} = h(0) = \int_{\phi=0} ds = |K|, \end{aligned}$$

which concludes the proof.

In general, this convergence result is valid for any approximations H_ε, δ_ε, under the assumptions

$$\lim_{\varepsilon\to 0} H_\varepsilon(z) = H(z) \text{ in } \mathbb{R}/\{0\},$$

$\delta_\varepsilon = H_\varepsilon{}'$, $H_\int \in C^1(\mathbb{R})$, $\int_{-\infty}^{+\infty} \delta_1(x) dx = 1$.

10 Piecewise-Constant Mumford and Shah Segmentation Using Level Sets

Our first formulation is for the case when the unknown set of edges K can be represented by $K = \{x \in \Omega: \phi(x) = 0\}$ for some (unknown) Lipschitz function $\phi:\Omega \to \mathbb{R}$. In this case, the unknown minimizers u to functions are restricted taking two unknown values c_1, c_2, and the corresponding Mumford-Shah minimization problem can be expressed as [29, 31]

$$\begin{aligned} \inf_{c_1,c_2,\phi} E(c_1, c_2, \phi) = &\int_\Omega (g(x) - c_1)^2 H(\phi) dx + \int_\Omega (g(x) - c_2)^2 H(-\phi) dx \\ &+ \nu_0 \int_\Omega |DH(\phi)|. \end{aligned} \tag{12}$$

The unknown minimizer u is expressed as

$$u(x) = c_1 H(\phi(x)) + c_2(1 - H(\phi(x))) = c_1 H(\phi(x)) + c_2 H(-\phi(x)).$$

H is substituted by its C^∞ approximation H_ε and instead

$$E_\varepsilon(c_1, c_2, \phi) = \int_\Omega (g(x) - c_1)^2 H_\varepsilon(\phi) dx + \int_\Omega (g(x) - c_2)^2 H_\varepsilon(-\phi) dx + v_0 \int_\Omega |\nabla H_\varepsilon(\phi)|\, dx. \tag{13}$$

is minimized.

The associated Euler-Lagrange equations with respect to c_1, c_2, and ϕ are

$$c_1(\phi) = \frac{\int_\Omega g(x) H_\varepsilon(\phi(x)) dx}{\int_\Omega H_\varepsilon(\phi(x)) dx}, c_1(\phi) = \frac{\int_\Omega g(x) H_\varepsilon(-\phi(x)) dx}{\int_\Omega H_\varepsilon(-\phi(x)) dx},$$

and, after simplifications,

$$\delta_\varepsilon(\phi) \left[(g(x) - c_1)^2 - (g(x) - c_2)^2 - \operatorname{div}\left(\frac{\nabla \phi}{|\nabla \phi|} \right) \right] = 0 \text{ in } \Omega, \tag{14}$$

with boundary conditions $\nabla \phi \cdot \vec{N} = 0$ on $\partial\Omega$. Since $\delta_\varepsilon > 0$ as defined, the factor $\delta_\varepsilon(\phi)$ can be removed from (14) or substituted by $|\nabla \phi|$ to obtain a more geometric motion extended to all level lines of ϕ, as in the standard level set approach.

This approach has been generalized by Chung and Vese in [34, 35], where more than one-level line of the same level set function ϕ can be used to represent the edge set K. Using m distinct real levels $\{l_1 < l_2 < \ldots < l_m\}$, the function ϕ partitions the domain Ω into the following $m+1$ disjoint open regions, making up Ω, together with their boundaries:

$$\begin{aligned} &\Omega_0 = \{x \in \Omega : -\infty < \phi(x) < l_1\}, \\ &\Omega_j = \{x \in \Omega : l_j < \phi(x) < l_{j+1}\}, \qquad 1 \le j \le m-1 \\ &\Omega_m = \{x \in \Omega : l_m < \phi(x) < +\infty\} \end{aligned}$$

The energy to minimize in this case, depending on $c_0, c_1, \ldots, c_m, \phi$, will be

$$\begin{aligned} E(c_0, c_1, \ldots, c_m, \phi) = &\int_\Omega |g(x) - c_0|^2 H(l_1 - \phi(x)) dx + \sum_{j=1}^{m-1} \int_\Omega |g(x) \\ &- c_j|^2 H(\phi(x) - l_j) H(l_{j+1} - \phi(x)) dx + \int_\Omega |g(x) \\ &- c_m|^2 H(\phi(x) - l_m) dx + v_0 \sum_{j=1}^{m} \int_\Omega |DH(\phi - l_j)|. \end{aligned} \tag{15}$$

The segmented image will be given by

$$\begin{aligned} u(x) = &c_0 H(l_1 - \phi(x)) + \sum_{j=1}^{m-1} c_j H(\phi(x) - l_j) H(l_{j+1} - \phi(x)) \\ &+ c_m H(\phi(x) - l_m). \end{aligned}$$

As before, to minimize the above energy, the Heaviside function H by H_ε, as $\varepsilon \to 0$, is approximated and substituted. The Euler-Lagrange equations associated with the corresponding minimization

$$\inf_{c_0,c_1,\ldots,c_m,\phi} E_\varepsilon(c_0, c_1, \ldots, c_m, \phi), \tag{16}$$

can be expressed as

$$\begin{cases} c_0(\phi) = \frac{\int_\Omega g(x) H_\varepsilon(l_1-\phi(t,x))\,dx}{\int_\Omega H_\varepsilon(l_1-\phi(t,x))dx}, \\ c_j(\phi) = \frac{\int_\Omega g(x) H_\varepsilon(\phi(t,x)-l_j) H_\varepsilon(l_{j+1}-\phi(t,x))dx}{\int_\Omega H_\varepsilon(\phi(t,x)-l_j) H_\varepsilon(l_{j+1}-\phi(t,x))dx}, \\ c_m(\phi) = \frac{\int_\Omega g(x) H_\varepsilon(\phi(t,x)-l_m)\,dx}{\int_\Omega H_\varepsilon(\phi(t,x)-l_m)dx}, \end{cases}$$

and

$$\begin{aligned} 0 = |g-c_0|^2 \delta_\varepsilon(l_1-\phi) + \textstyle\sum_{j=1}^{m-1} \left|g-c_j\right|^2 [\delta_\varepsilon(l_{j+1}-\phi) H_\varepsilon(\phi-l_j)-] \\ \delta_\varepsilon(\phi-l_j) H_\varepsilon(l_{j+1}-\phi) - |g-c_m|^2 \delta_\varepsilon(\phi-l_m)+ \\ v_0 \textstyle\sum_{j=1}^{m} \left[\delta_\varepsilon(\phi-l_j)\,\mathrm{div}\left(\frac{\nabla\phi}{|\nabla\phi|}\right)\right],\ \frac{\partial\phi}{\partial\vec{n}}\Big|_{\partial\Omega} = 0, \end{aligned}$$

where $\vec{N}$ is the exterior unit normal to the boundary $\partial\Omega$.

Given here are the details of the numerical algorithm for solving (17) in two dimensions (x, y), using gradient descent, in the case of one function ϕ with two levels $l_1 = 0, l_2 = l > 0$. Let $h = \Delta x = \Delta y$ be the space steps, Δt be the time step, and $\varepsilon = h$. Let (x_i, y_j) be the discrete points, for $1 \le i, j \le M$, and $g_{i,j} \approx g(x_i, y_j)$, $\phi^n_{i,j} \approx \phi(n\Delta t, x_i, y_j)$, with $n \ge 0$. Recall the usual finite differences formulas

$$\begin{aligned} \Delta^x_+\phi_{i,j} &= \phi_{i+1,j} - \phi_{i,j}, \quad \Delta^x_-\phi_{i,j} = \phi_{i,j} - \phi_{i-1,j}, \\ \Delta^y_+\phi_{i,j} &= \phi_{i,j+1} - \phi_{i,j}, \quad \Delta^x_-\phi_{i,j} = \phi_{i,j} - \phi_{i,j-1} \end{aligned}$$

Set $n = 0$, and start with $\phi^0_{i,j}$ given (defining the initial set of curves). Then, for each $n > 0$ until steady state:

1. Compute averages $c_0(\phi^n)$, $c_1(\phi^n)$, and $c_2(\phi^n)$.
2. Compute $\phi^{n+1}_{i,j}$, derived from the finite differences scheme:

$$\begin{aligned} \frac{\phi^{n+1}_{i,j}-\phi^n_{i,j}}{\Delta t} = \delta_\varepsilon\left(\phi^n_{i,j}\right)\left[\frac{v_0}{h^2}\left(\Delta^x_-\left(\frac{\phi^n_{i+1,j}-\phi^{n+1}_{i,j}}{\left|\nabla\phi^n_{i,j}\right|}\right)+\Delta^y_-\left(\frac{\phi^n_{i,j+1}-\phi^{n+1}_{i,j}}{\left|\nabla\phi^n_{i,j}\right|}\right)\right)\right. \\ \left.+\left|g_{i,j}-c_0\right|^2-\left|g_{i,j}-c_1\right|^2 H_\varepsilon\left(l-\phi^n_{i,j}\right)\right]+\delta_\varepsilon\left(\phi^n_{i,j}-l\right) \\ \left[\frac{v_0}{h^2}\left(\Delta^x_-\left(\frac{\phi^n_{i+1,j}-\phi^{n+1}_{i,j}}{\left|\nabla\phi^n_{i,j}\right|}\right)+\Delta^y_-\left(\frac{\phi^n_{i,j+1}-\phi^{n+1}_{i,j}}{\left|\nabla\phi^n_{i,j}\right|}\right)\right)\right. \\ \left.-\left|g_{i,j}-c_2\right|^2+\left|g_{i,j}-c_1\right|^2 H_\varepsilon\left(\phi^n_{i,j}\right)\right], \end{aligned}$$

3. Where $\left|\nabla\phi_{i,j}^n\right| = \sqrt{\left(\frac{\phi_{i+1,j}^n-\phi_{i,j}^n}{h}\right)^2+\left(\frac{\phi_{i,j+1}^n-\phi_{i,j}^n}{h}\right)^2}$. Let

$$C_1 = \frac{1}{\sqrt{\left(\frac{\phi_{i+1,j}^n-\phi_{i,j}^n}{h}\right)^2+\left(\frac{\phi_{i,j+1}^n-\phi_{i,j}^n}{h}\right)^2}},$$
$$C_2 = \frac{1}{\sqrt{\left(\frac{\phi_{i,j}^n-\phi_{i-1,j}^n}{h}\right)^2+\left(\frac{\phi_{i-1,j+1}^n-\phi_{i-1,j}^n}{h}\right)^2}},$$
$$C_3 = \frac{1}{\sqrt{\left(\frac{\phi_{i+1,j}^n-\phi_{i,j}^n}{h}\right)^2+\left(\frac{\phi_{i,j+1}^n-\phi_{i,j}^n}{h}\right)^2}},$$
$$C_4 = \frac{1}{\sqrt{\left(\frac{\phi_{i+1,j-1}^n-\phi_{i,j-1}^n}{h}\right)^2+\left(\frac{\phi_{i,j}^n-\phi_{i,j-1}^n}{h}\right)^2}}$$

4. Let $m_1 = \frac{\Delta t}{h^2}\left(\delta_\varepsilon\left(\phi_{i,j}^n\right)+\delta_\varepsilon\left(\phi_{i,j}^n-l\right)\right)v_0$, $C = 1 + m_1(C_1 + C_2 + C_3 + C_4)$. The main update equation for ϕ becomes

$$\phi_{i,j}^{n+1} = \frac{1}{C}\left[\phi_{i,j}^n + m_1\left(C_1\phi_{i+1,j}^n + C_2\phi_{i-1,j}^n + C_3\phi_{i,j+1}^n + C_4\phi_{i,j-1}^n\right)\right.$$
$$+\Delta t\delta_\varepsilon\left(\phi_{i,j}^n\right)(-(g_{i,j}-c_1)^2\left(1-H_\varepsilon\left(\phi_{i,j}^n-l\right)\right)$$
$$\left.+(g_{i,j}-c_0)^2) + \Delta t\delta_\varepsilon\left(\phi_{i,j}^n-l\right)\left(-(g_{i,j}-c_2)^2+(g_{i,j}-c_1)^2H_\varepsilon\left(\phi_{i,j}^n\right)\right)\right],$$

5. And repeat, until steady state is reached.

This section is concluded with several experimental results obtained using the models presented here that act as denoising, segmentation, and active contours. In Fig. 4, an experimental result is shown taken from [31] obtained using the binary piecewise-constant model (12); we notice how interior contours can be automatically detected. In Fig. 5, we show an experimental result using the multilayer model (15), with $m = 2$ and two levels l_1, l_2, applied to the segmentation of a brain image.

The work in [35, 79] also shows how the previous Mumford-Shah level set approaches can be extended to piecewise-constant segmentation of images with triple junctions, several non-nested regions, or with other complex topologies, by using two or more level set functions that form a perfect partition of the domain Ω.

11 Piecewise-Smooth Mumford and Shah Segmentation Using Level Sets

Considered first is the corresponding two-dimensional case under the assumption that the edges denoted by K in the image can be represented by one-level set function ϕ, i.e., $K = \{x \in \Omega | \phi(x) = 0\}$, and followed are the approaches developed in parallel by Chan and Vese [32, 79] and by Tsai et al. [78], in order to minimize

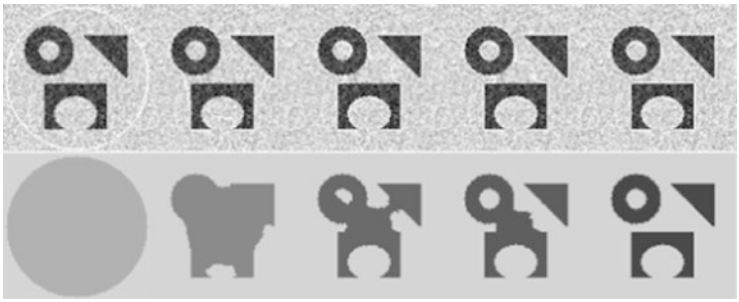


Fig. 4 Detection of different objects in a noisy image, with various convexities and with an interior contour which is automatically detected, using only one initial curve. After a short time, an interior contour appears inside the torus, and then it expands. *Top*: g and the evolving contours. *Bottom*: the piecewise-constant approximations u of g over time, given by $u = c_1 H(\phi) + c_2(1 - H(\phi))$

the general Mumford and Shah model. As in [79], the link between the unknowns u and ϕ can be expressed by introducing two functions u^+ and u^- (see Fig. 6) such that

$$u(x) = \begin{cases} u^+(x) \text{ if } \phi(x) \geq 0, \\ u^-(x) \text{ if } \phi(x) \leq 0. \end{cases}$$

Assume that u^+ and u^- are H^1 functions on $\phi \geq 0$ and on $\phi \leq 0$, respectively (with Sobolev traces up to all boundary points, i.e., up to the boundary $\{\phi = 0\}$). The following minimization problem can be written

$$\inf_{u^+,u^-,\phi} E(u^+, u^-, \phi),$$

where

$$E(u^+, u^-, \phi) = \mu^2 \int_\Omega \left|u^+ - g\right|^2 H(\phi)dx + \mu^2 \int_\Omega |u^- - g|^2 (1 - H(\phi))dx \\ + \int_\Omega \left|\nabla u^+\right|^2 H(\phi))dx + \int_\Omega |\nabla u^-|^2 (1 - H(\phi))dx + \nu \int_\Omega |DH(\phi)|$$

is the Mumford-Shah functional restricted to $u(x) = u^+(x)H(\phi(x)) + u^-(x)(1 - H(\phi(x))$. Minimizing $E(u^+, u^-, \phi)$ with respect to u^+, u^-, and ϕ, we obtain the following Euler-Lagrange equations (embedded in a time-dependent dynamical scheme for ϕ):

$$\mu^2(u^+ - g) = \Delta u^+ \text{in } \{x : \phi(t, x) > 0\}, \frac{\partial u^+}{\partial \vec{n}} = 0 \text{ on } \{x : \phi(t, x) = 0\} \cup \partial\Omega, \tag{17}$$

$$\mu^2(u^- - g) = \Delta u^- \text{in } \{x : \phi(t, x) < 0\}, \frac{\partial u^-}{\partial \vec{n}} = 0 \text{ on } \{x : \phi(t, x) = 0\} \cup \partial\Omega, \tag{18}$$

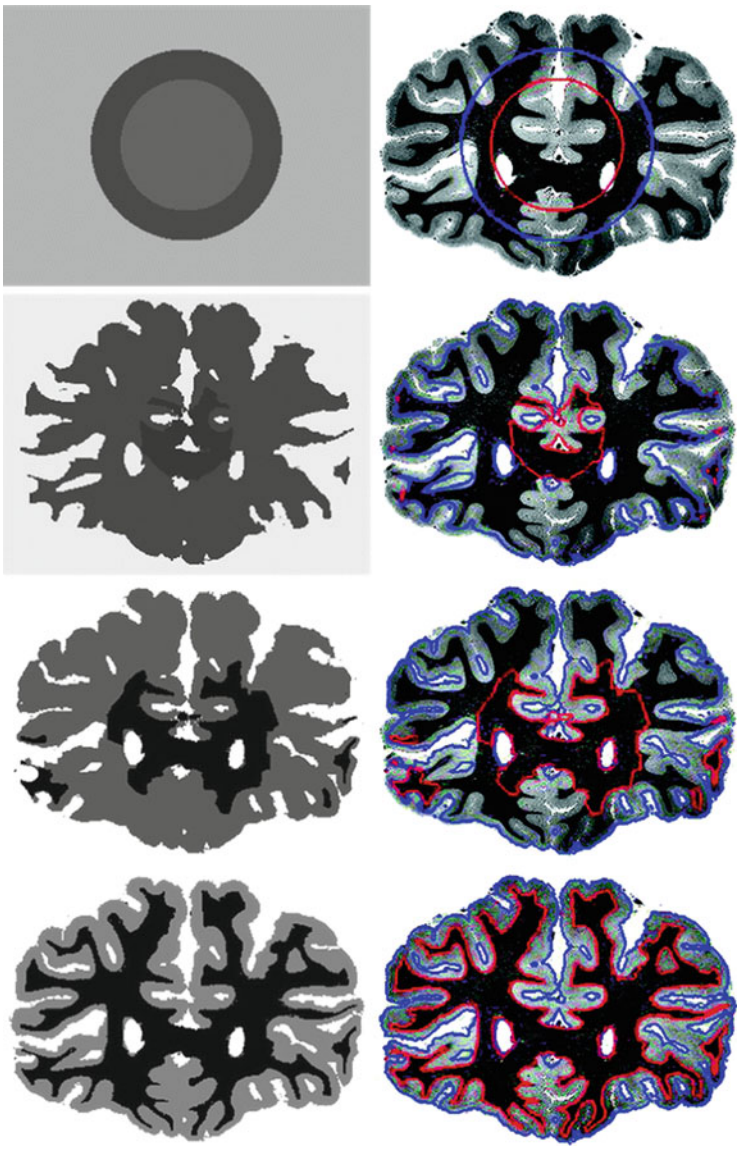


Fig. 5 Segmentation of a brain image using one-level set function with two levels. Parameters: $l_1 = 0, l_2 = 25, \Delta t = 0.1, \nu_0 = 0.1 \cdot 255^2$, 1,500 iterations

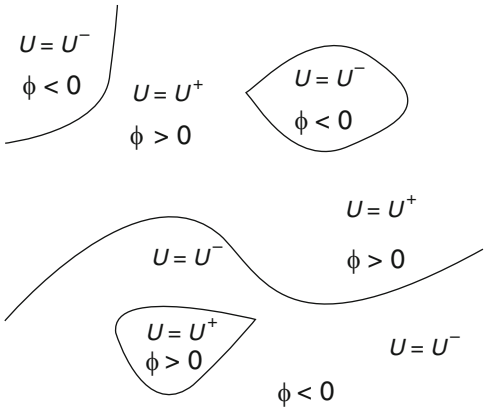


Fig. 6 The functions u^+, u^- and the zero-level lines of the level set function ϕ for piecewise-smooth image partition

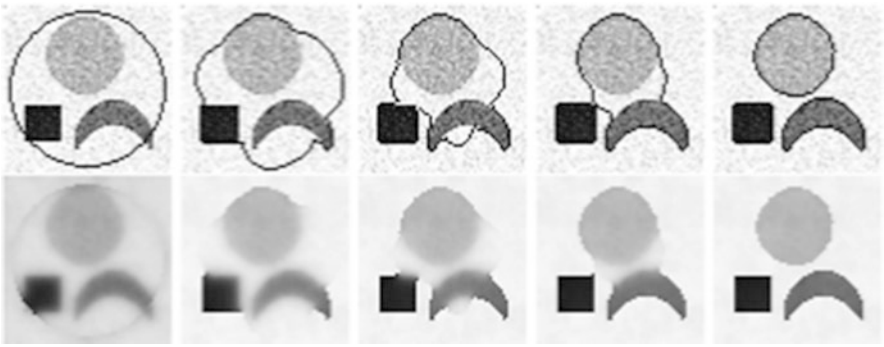


Fig. 7 Results on a noisy image, using the level set algorithm for the piecewise-smooth Mumford-Shah model with one-level set function. The algorithm performs as active contours, denoising, and edge detection

$$\frac{\partial \phi}{\partial t} = \delta_\varepsilon(\phi)\left[\nu \nabla\left(\frac{\nabla \phi}{|\nabla \phi|}\right) - \mu^2 \left|u^+ - g\right|^2 - \left|\nabla u^+\right|^2 + \mu^2 \left|u^- - g\right|^2 + \left|\nabla u^-\right|^2\right], \tag{19}$$

where $\partial/\partial \vec{n}$ denotes the partial derivative in the normal direction $\vec{n}$ at the corresponding boundary. We also associate the boundary condition $\frac{\partial \phi}{\partial \vec{n}} = 0$ on $\partial\Omega$ to Eq. (19).

Shown in Figs. 7 and 8 are experimental results taken from [79] obtained with the piecewise-smooth two-phase model.

There are cases when the boundaries K of regions forming a partition of the image could not be represented by the boundary of an open domain. To overcome this, several solutions have been proposed in this framework, and mentioned here are two of them: (1) in the work by Tsai et al. [78], the minimization of $E(u^+, u^-, \phi)$ is repeated inside each of the two regions previously computed, and (2) in the work of Chan and Vese [79], two or more level set functions are used. For example, in

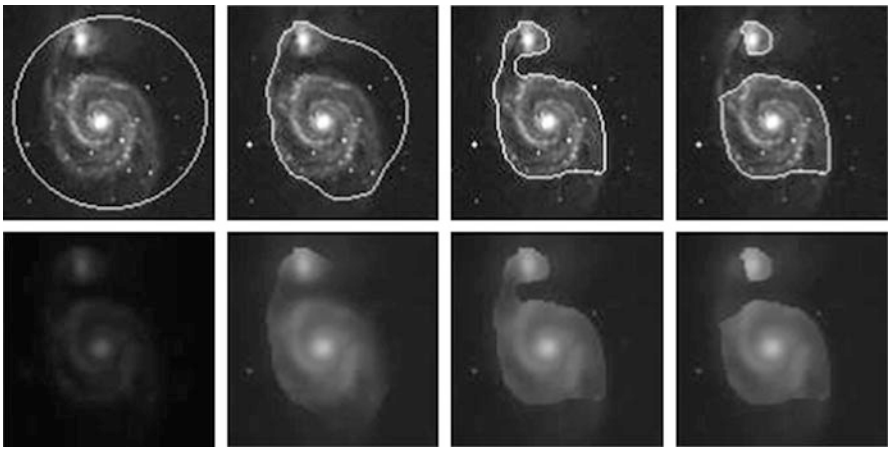


Fig. 8 Numerical result using the piecewise-smooth Mumford-Shah level set algorithm with one-level set function, on a piecewise-smooth real galaxy image

two dimensions, the problem can be solved using only two-level set functions, and we do not have to know a priori how many gray levels the image has (or how many segments). The idea is based on the four-color theorem. Based on this observation, we can "color" all the regions in a partition using only four "colors," such that any two adjacent regions have different "colors." Therefore, using two-level set functions, we can identify the four "colors" by the following (disjoint) sets: $\{\phi_1 > 0, \phi_2 > 0\}$, $\{\phi_1 < 0, \phi_2 < 0\}$, $\{\phi_1 < 0, \phi_2 > 0\}$, $\{\phi_1 < 0, \phi_2 < 0\}$. The boundaries of the regions forming the partition will be given by $\{\phi_1 = 0\} \cup \{\phi_2 = 0\}$, and this will be the set of curves K. Note that, in this particular multiphase formulation of the problem, we do not have the problems of "overlapping" or "vacuum" (i.e., the phases are disjoint, and their union is the entire domain Ω, by definition).

As before, the link between the function u and the four regions can be made by introducing four functions u^{++}, u^{+-}, u^{-+}, u^{--}, which are in fact the restrictions of u to each of the four phases, as follows (see Fig. 9):

$$u(x) = \begin{cases} u^{++}(x), & \text{if } \phi_1(x) > 0 \text{ and } \phi_2(x) > 0, \\ u^{+-}(x), & \text{if } \phi_1(x) > 0 \text{ and } \phi_2(x) < 0, \\ u^{-+}(x), & \text{if } \phi_1(x) < 0 \text{ and } \phi_2(x) > 0, \\ u^{--}(x), & \text{if } \phi_1(x) < 0 \text{ and } \phi_2(x) < 0. \end{cases}$$

Again, using the Heaviside function, the relation between u; the four functions u^{++}, u^{+-}, u^{-+}, u^{--}; and the level set functions ϕ_1 and ϕ_2 can be expressed by

$$u = u^{++}H(\phi_1)H(\phi_2) + u^{+-}H(\phi_1)(1 - H(\phi_2)) + u^{-+}(1 - H(\phi_1))H(\phi_2) + u^{--}(1 - H(\phi_1))(1 - H(\phi_2))$$

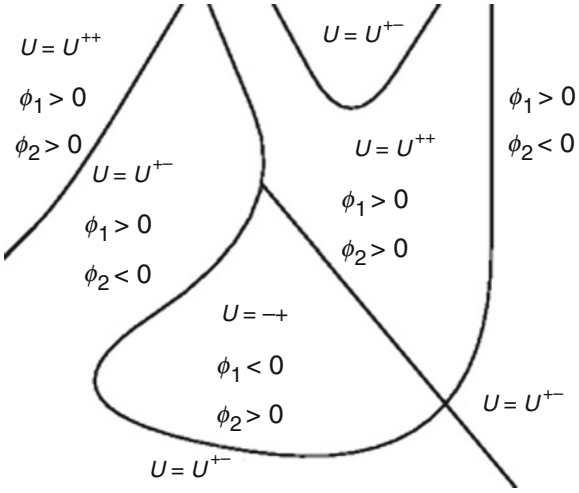


Fig. 9 The functions u^{++}, u^{+-}, u^{-+}, u^{--}, and the zero-level lines of the level set functions ϕ_1, ϕ_2 for piecewise-smooth image partition

Introduced here is an energy in level set formulation based on the Mumford-Shah functional:

$$\begin{aligned}
E(u,\phi_1,\phi_2) = {} & \mu^2 \int_\Omega \left|u^{++}-g\right|^2 H(\phi_1)H(\phi_2)dx \\
& + \int_\Omega \left|\nabla u^{++}\right|^2 H(\phi_1)H(\phi_2)\,dx \\
& + \mu^2 \int_\Omega \left|u^{+-}-g\right|^2 H(\phi_1)(1-H(\phi_2))\,dx \\
& + \int_\Omega \left|\nabla u^{+-}\right|^2 H(\phi_1)(1-H(\phi_2))\,dx \\
& + \mu^2 \int_\Omega \left|u^{-+}-g\right|^2 (1-H(\phi_1))H(\phi_2)\,dx \\
& + \int_\Omega \left|\nabla u^{-+}\right|^2 (1-H(\phi_1)))H(\phi_2)dx \\
& + \mu^2 \int_\Omega |u^{--}-g|^2 (1-H(\phi_1))(1-H(\phi_2))dx \\
& + \int_\Omega |\nabla u^{--}|^2 (1-H(\phi_1))(1-H(\phi_2))dx \\
& + \nu \int_\Omega |DH(\phi_1)| + \nu \int_\Omega |DH(\phi_2)|
\end{aligned}$$

Note that the expression $\int_\Omega |DH(\phi_1)| + \int_\Omega |DH(\phi_2)|$ is not exactly the length term of $K = \{x \in \Omega\colon \phi_1(x) = 0 \text{ and } \phi_2(x) = 0\}$; it is just an approximation and simplification. In practice, satisfactory results using the above formula are obtained, and the associated Euler-Lagrange equations are simplified.

The associated Euler-Lagrange equations are obtained as in the previous cases, embedded in a dynamic scheme, assuming $(t, x, y) \mapsto\!\!\to \phi_i\ (t, x, y)$: minimizing the energy with respect to the functions u^{++}, u^{+-}, u^{-+}, u^{--}, we have, for each fixed t:

$$\mu^2(u^{++}-g) = \Delta u^{++} \text{ in } \{\phi_1 > 0, \phi_2 > 0\}, \tfrac{\partial u^{++}}{\partial \vec{n}} = 0 \text{ on } \{\phi_1 = 0,\ \phi_2 \geq 0\},$$
$$\{\phi_1 \geq 0,\ \phi_2 = 0\};$$
$$\mu^2(u^{+-}-g) = \Delta u^{+-} \text{ in } \{\phi_1 > 0, \phi_2 < 0\}, \tfrac{\partial u^{+-}}{\partial \vec{n}} = 0 \text{ on } \{\phi_1 = 0,\ \phi_2 \leq 0\},$$

$$\{\phi_1 \geq 0,\ \phi_2 = 0\};$$

$$\mu^2(u^{-+} - g) = \Delta u^{-+} \text{ in } \{\phi_1 < 0, \phi_2 > 0\},\ \frac{\partial u^{-+}}{\partial \vec{n}} = 0 \text{ on } \{\phi_1 = 0,\ \phi_2 \geq 0\},$$

$$\{\phi_1 \leq 0,\ \phi_2 = 0\};$$

$$\mu^2(u^{--} - g) = \Delta u^{--} \text{ in } \{\phi_1 < 0, \phi_2 < 0\},\ \frac{\partial u^{--}}{\partial \vec{n}} = 0 \text{ on } \{\phi_1 = 0,\ \phi_2 \leq 0\},$$

$$\{\phi_1 \leq 0,\ \phi_2 = 0\}$$

The Euler-Lagrange equations evolving ϕ_1 and ϕ_2, embedded in a dynamic scheme, are formally:

$$\begin{aligned}
\frac{\partial \phi_1}{\partial t} &= \delta_\varepsilon(\phi_1)\left[\nu\nabla\left(\frac{\nabla\phi_1}{|\nabla\phi_1|}\right) - \mu^2\left|u^{++} - g\right|^2 H(\phi_2) - \left|\nabla u^{++}\right|^2 H(\phi_2)\right. \\
&\quad -\mu^2\left|u^{+-} - g\right|^2 (1 - H(\phi_2)) - \left|\nabla u^{+-}\right|^2 (1 - H(\phi_2)) + \mu^2\left|u^{-+}\right. \\
&\quad \left.-g|^2 H(\phi_2) + \left|\nabla u^{-+}\right|^2 H(\phi_2) + \mu^2 |u^{--} - g|^2 (1 - H(\phi_2)) + \right. \\
&\quad \left.|\nabla u^{--}|^2 (1 - H(\phi_2))\right] = 0, \\
\frac{\partial \phi_2}{\partial t} &= \delta_\varepsilon(\phi_2)\left[\nu\nabla\left(\frac{\nabla\phi_2}{|\nabla\phi_2|}\right) - \mu^2\left|u^{++} - g\right|^2 H(\phi_1) - \left|\nabla u^{++}\right|^2 H(\phi_1)\right. \\
&\quad -\mu^2\left|u^{+-} - g\right|^2 H(\phi_2) + \left|\nabla u^{+-}\right|^2 H(\phi_1) - \mu^2\left|u^{-+} - g\right|^2 (1 - H(\phi_1)) - \\
&\quad \left|\nabla u^{-+}\right|^2 (1 - H(\phi_1)) \\
&\quad \left.-\mu^2 |u^{--} - g|^2 (1 - H(\phi_1)) + |\nabla u^{--}|^2 (1 - H(\phi_1))\right]
\end{aligned}$$

It can be shown, by standard techniques of the calculus of variations on the space $\mathrm{SBV}(\Omega)$ (special functions of bounded variations), and a compactness result due to Ambrosio [3], that the proposed minimization problems from this section, in the level set formulation, have a minimizer. Finally, because there is no uniqueness of minimizers, and because the problems are nonconvex, the numerical results may depend on the initial choice of the curves and may compute only a local minimum. We think that, using the seed initialization (see [79]), the algorithms have the tendency of computing a global minimum, most of the times. Additional experimental results are shown in [79].

12 Case Examples: Variational Image Restoration with Segmentation-Based Regularization

This section focuses on the challenging task of edge-preserving variational image restoration. In this context, restoration is referred to as image deblurring and denoising, dealing with Gaussian and impulsive noise models. Terms from the Mumford-Shah segmentation functional are used as regularizers, reflecting the model of piecewise-constant or piecewise-smooth images.

In the standard model of degradation, the underlying assumptions are the linearity and shift invariance of the blur process and the additivity and normal

distribution of the noise. Formally, let Ω be an open-bounded subset of $\mathbb{R}^n$. The observed image g: $\Omega \to \mathbb{R}^N \in L^\infty$ is given by

$$g = h * u + n, \tag{20}$$

where g is normalized to the hypercube $[0, 1]^N$, h is the blur kernel such that $h(x) > 0$ and $\int h(x)dx = 1$, u: $\Omega \to \mathbb{R}^N$ is the ("ideal") original image, $n \sim N(0, \sigma^2)$ stands for a white Gaussian noise, and * denotes the convolution operator. The restoration problem is the recovery of the original image u given Eq. (20). Non-blind image restoration is the problem whenever the blur kernel is known, while blind restoration refers to the case of unknown kernel [49,50]. The recovery process in the non-blind case is a typical inverse problem where the image u is the minimizer of an objective functional of the form

$$F(u) = \Phi(g - h * u) + J(\nabla u). \tag{21}$$

The functional consists of fidelity term and a regularizer. The fidelity term ϕ forces the smoothed image $h * u$ to be close to the observed image g. The commonly used model of a white Gaussian noise $n \sim N(0, \sigma^2)$ leads by the maximum likelihood estimation to the minimization of the L^2 norm of the noise

$$\Phi_{L^2} = \|g - h * u\|^2_{L^2(\Omega)}. \tag{22}$$

However, in the case of impulsive noise, some amount of pixels do not obey the Gaussian noise model. Minimization of outlier effects can be accomplished by replacing the quadratic form (22) with a robust ρ-function [44], e.g.,

$$\Phi_{L^1} = \|g - h * u\|_{L^1(\Omega)}. \tag{23}$$

The minimization of (22) or (23) alone with respect to u is an inverse problem which is known to be ill-posed: small perturbations in the data g may produce unbounded variations in the solution. To alleviate this problem, a regularization term can be added. The Tikhonov L^2 stabilizer [77]

$$J_{L^2} = \int_\Omega |\nabla u|^2 \, dx,$$

leads to over smoothing and loss of important edge information. A better edge preservation regularizer, the total variation (TV) term, was introduced by Rudin et al. [69,70], where the L^2 norm was replaced by the L^1 norm of the image gradients

$$J_{L^1} = \int_\Omega |\nabla u| \, dx.$$

Still, although the total variation regularization outperforms the L^2 norm, the image features – the edges – are not explicitly extracted. The edges are implicitly preserved only by the image gradients.

An alternative regularizer is the one used in the Mumford-Shah functional [61, 63]. Recall that this is accomplished by searching for a pair (u, K) where $K \subset \Omega$ denotes the set of discontinuities of u, the unknown image, such that $u \in H^1(\Omega \backslash K)$, $K \subset \Omega$ closed in Ω, and

$$G(u, K) = \beta \int_{\Omega/K} |\nabla u|^2 \, dx + \alpha H^{n-1}(K) < \infty. \tag{24}$$

In our study, the regularizer to the restoration problem (21) is given by

$$J_{\mathrm{MS}} = G(u, K),$$

its L^1 variant [2, 74], and elliptic or level set approximations of these, as presented next. This enables the explicit extraction and preservation of the image edges in the course of the restoration process. We show the advantages of this regularizer in several applications and noise models (Gaussian and impulsive).

As has been mentioned, Ambrosio and Tortorelli [6] introduced an elliptic approximation $G_\varepsilon(u, v)$ to $G(u, K)$, as $\varepsilon \to 0^+$, that is recalled here,

$$G_\varepsilon(u, v) = \beta \int_\Omega v^2 |\nabla u|^2 \, dx + \alpha \int_\Omega \left(\varepsilon |\nabla v|^2 + \frac{(v-1)^2}{4\varepsilon} \right) dx. \tag{25}$$

Replacing the Mumford-Shah regularization term (24) by $G_\varepsilon(u, v)$ yields the proposed restoration model

$$F_\varepsilon(u, v) = \Phi(g - h * u) + \beta \int_\Omega v^2 |\nabla u|^2 \, dx + \alpha \int_\Omega \left(\varepsilon |\nabla v|^2 + \frac{(v-1)^2}{4\varepsilon} \right) \tag{26}$$

The functional (26) can also be understood from a generalized robust statistics viewpoint. This is beyond the scope of this chapter, and the interested reader can find the details in [12].

The rest of the chapter considers the non-blind restoration problem presented in [13] and its generalizations to several more realistic situations. Consider the problem of (semi-) blind deconvolution, the case of impulsive noise, the color restoration problem, and the case of space-variant blur. Also consider the problem of restoration of piecewise-constant images from noisy-blurry data using the level set form of the Mumford-Shah regularizer and image restoration using nonlocal Mumford-Shah-Ambrosio-Tortorelli regularizers.

13 Non-blind Restoration

Addressed first is the restoration problem with a known blur kernel h and additive Gaussian noise [10, 13]. In this case, the fidelity term is the L^2 norm of the noise (22), and the regularizer $\vartheta_{\mathrm{MS}} = G_f\ (u,v)$ (25). The objective functional is therefore

$$F_\varepsilon(u,v) = \frac{1}{2}\int_\Omega (g - h * u)^2\, dx + \beta \int_\Omega v^2\, |\nabla u|^2\, dx + \alpha \int_\Omega \left(\varepsilon\, |\nabla v|^2 + \frac{(v-1)^2}{4\varepsilon}\right) \tag{27}$$

The functional (27) is strictly convex, bounded from below and coercive with respect to the functions u and v if the other one is fixed. Following [33], the alternate minimization (AM) approach is applied: in each step of the iterative procedure, we minimize with respect to one function and keep the other one fixed. The minimization is carried out using the Euler-Lagrange (E-L) equations with Neumann boundary conditions where u is initialized as the blurred image g and v is initialized to 1.

$$\frac{\delta F_\varepsilon}{\delta v} = 2\beta v\, |\nabla u|^2 + \alpha \frac{v-1}{2\varepsilon} - 2\varepsilon\alpha\, \Delta\, v = 0 \tag{28}$$

$$\frac{\delta F_\varepsilon}{\delta u} = (h * u - g) * h\,(-x,-y) - 2\beta \nabla \cdot \left(v^2 \nabla u\right) = 0 \tag{29}$$

Equation (28) is linear with respect to v and can be easily solved after discretization by the minimal residual algorithm [81]. The integrodifferential equation (29) can be solved by the conjugate-gradient method [13]. The iterative process is stopped whenever some convergence criterion is satisfied (e.g., $||u^{n+1} - u^n|| < \varepsilon||u^n||$). Figure 10 demonstrates the outcome of the algorithm. The top-left image is the blurred image g. The kernel corresponds to horizontal motion blur. The top-right image is the reconstruction obtained using total variation (TV) regularization [70, 80]. The bottom-left image is the outcome of the MS regularizer, with a known blur kernel. The bottom-right image shows the associated edge map v determined by the algorithm. Acceptable restoration is obtained with both methods. Nevertheless, the MS method yields a sharper result and is almost free of "ghosts" (white replications of notes) that can be seen in the top-right image (e.g., between the C notes in the right part of the top stave). The algorithm can be also applied to 3D images as shown in Fig. 11. In this example, the blur kernel was anisotropic 3D Gaussian kernel.

14 Semi-blind Restoration

Blind restoration refers to the case when the blur kernel h is not known in advance. In addition to being ill-posed with respect to the image, the blind restoration problem is ill-posed in the kernel as well. Blind image restoration with joint recovery of the image and the kernel, and regularization of both, was presented by You and Kaveh

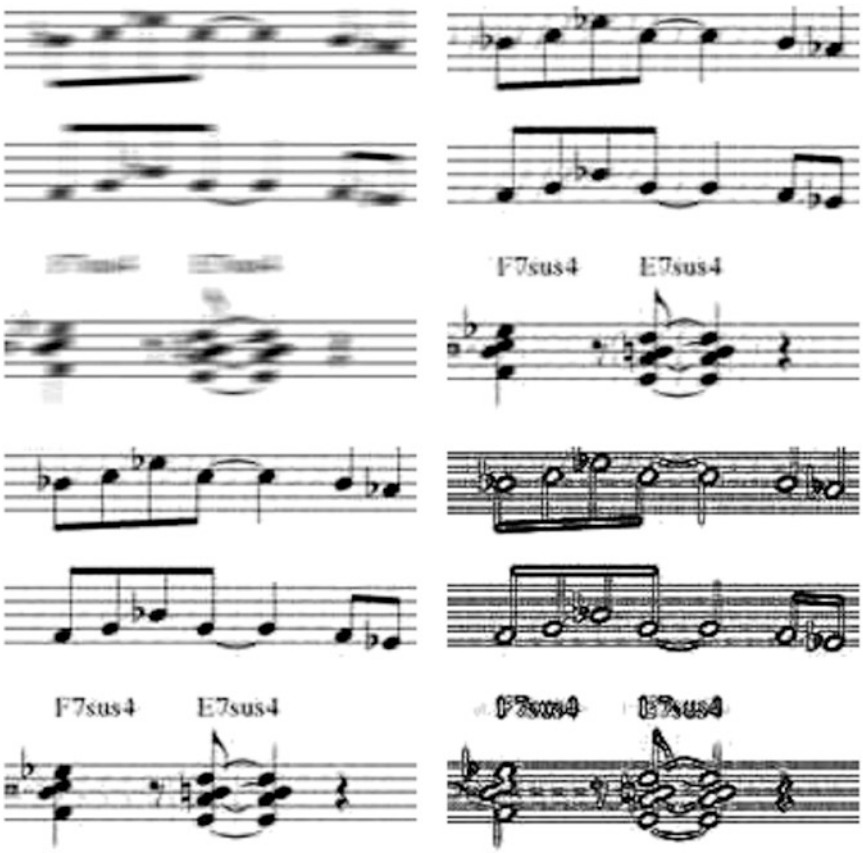



Fig. 10 The case of a known (nine-pixel horizontal motion) blur kernel. *Top left*: corrupted image. *Top-right*: restoration using the TV method [70, 80]. *Bottom left*: restoration using the MS method. *Bottom right*: edge map produced by the MS method

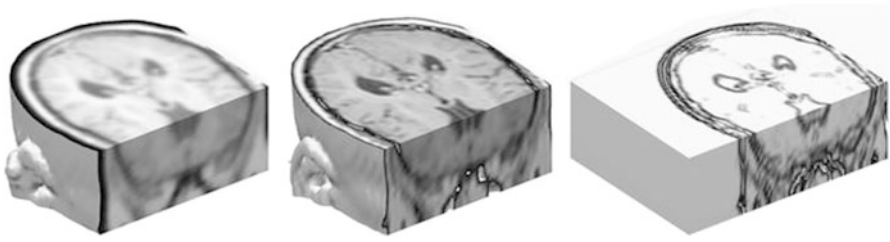


Fig. 11 3D restoration of MR image. *Left*: blurred ($\sigma_x = 1.0, \sigma_y = 1.0, \sigma_z = 0.2$) image. *Middle*: recovered image. *Right*: edge map

[82], followed by Chan and Wong [33]. Chan and Wong suggested to minimize a functional consisting of a fidelity term and total variation (L^1 norm) regularization for both the image and the kernel:

$$F(u,h) = \frac{1}{2}\|h * u - g\|^2_{L^2(\Omega)} + \alpha_1 \int_\Omega |\nabla u|\, dx + \alpha_2 \int_\Omega |\nabla h|\, dx. \tag{30}$$

By this approach, the recovered kernel is highly dependent on the image characteristics. It allows the distribution of edge directions in the image to have an influence on the shape of the recovered kernel which may lead to inaccurate restoration [13]. Facing the ill-posedness of blind restoration with a general kernel, two approaches can be taken. One is to add relevant data; the other is to constrain the solution. In many practical situations, the blurring kernel can be modeled by the physics/optics of the imaging device and the setup. The blurring kernel can then be constrained and described as a member in a class of parametric functions. The blind restoration problem is then reduced to a semi-blind one. Let us consider the case of isotropic Gaussian blur parameterized by the width σ,

$$h_\sigma(x) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2}{2\sigma^2}}, x \in \mathbb{R}^n.$$

The semi-blind objective functional then takes the form [13]

$$F_\varepsilon(v,u,\sigma) = \frac{1}{2}\int_\Omega (h_\sigma * u - g)^2\, dx + G_\varepsilon(u,v) + \gamma \int_\Omega |\nabla h_\sigma|^2\, dx. \tag{31}$$

The last term in Eq. (31) stands for the regularization of the kernel, necessary to resolve the fundamental ambiguity in the division of the apparent blur between the recovered image and the blur kernel. This means that we prefer to reject the hypothesis that the blur originates from u, and assume that it is due to the convolution with the blur kernel. From the range of possible kernels, we thus select a wide one. This preference is represented by the kernel smoothness term: the width of the Gaussian corresponds to its smoothness, measured by the L^2 norm of its gradient. The optimization is carried out by using the alternate minimization approach. The recovered image u is initialized with g, the edge indicator function v is initialized with 1, and σ with a small number ε which reflects a delta function kernel. The Euler-Lagrange equations with respect to v and u are given by (28) and (29), respectively. The parameter σ is the solution of

$$\frac{\partial F_\varepsilon}{\partial \sigma} = \int_\Omega \left[(h_\sigma * u - g)\left(\frac{\partial h_\sigma}{\partial \sigma} * u\right) + \gamma \frac{\partial}{\partial \sigma} |\nabla h_\sigma|^2 \right] dx = 0, \tag{32}$$

which can be calculated by the bisection method. The functional (31) is not generally convex. Nevertheless, in practical numerical simulations, the algorithm converges to visually appealing restoration results as can be seen in the second row of Fig. 12.

Fig. 12 Semi-blind restoration. *Top row*: blurred images. *Second row*: restoration using the semi-blind method. *Third row*: original images. *Bottom row*: edge maps produced by the semi-blind method

15 Image Restoration with Impulsive Noise

Consider an image that has been blurred with a known blur kernel h and contaminated by impulsive noise. Salt-and-pepper noise, for instance, is a common model for the effects of bit errors in transmission, malfunctioning pixels, and faulty memory locations. Image deblurring algorithms that were designed for Gaussian noise produce inadequate results with impulsive noise.

The left image of Fig. 13 is a blurred image contaminated by salt-and-pepper noise, and the right image is the outcome of the total variation restoration method [80]. A straightforward sequential approach is to first denoise the image and then to deblur it. This two-stage method is however prone to failure, especially at high

Fig. 13 Current image deblurring algorithms fail in the presence of salt-and-pepper noise. *Left*: blurred image with salt-and-pepper noise. *Right*: restoration using the TV method [80]

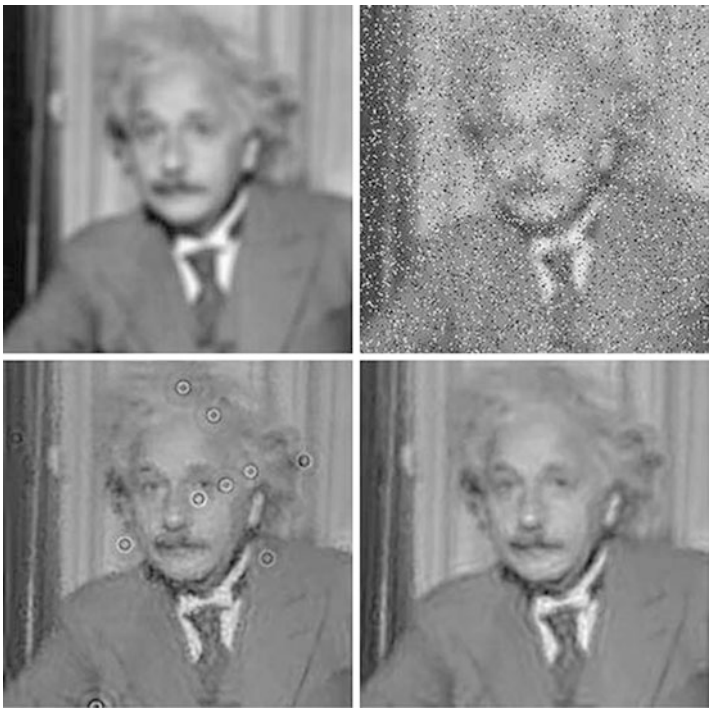


Fig. 14 The failure of the two-stage approach to salt-and-pepper noise removal and image deblurring. *Top-left*: blurred image. *Top-right*: blurred image contaminated by salt-and-pepper noise. *Bottom-left*: the outcome of 3×3 median filtering, followed by deblurring. *Bottom-right*: the outcome of 5×5 median filtering, followed by deblurring

noise density. Image denoising using median-type filtering creates distortion that depends on the neighborhood size; this error can be strongly amplified by the deblurring process. This is illustrated in Fig. 14. The top-left and top-right images are the blurred and blurred-noisy images, respectively. The outcome of 3×3 median filtering followed by total variation deblurring [80] is shown in the bottom left. At this noise level, the 3×3 neighborhood size of the median filter is insufficient, the noise is not entirely removed, and the residual noise is greatly amplified by the deblurring process. If the neighborhood size of the median filter increases to 5×5, the noise is fully removed, but the distortion leads to inadequate deblurring (bottom right).

In a unified variational framework, the "ideal" image u can be approximated as the minimizer of the objective functional [12, 14]

$$F_\varepsilon(u, v) = \int_\Omega \sqrt{(h * u - g)^2 + \eta}\, dx + G_\varepsilon(u, v). \tag{33}$$

The quadratic data-fidelity term is now replaced by the modified L^1 norm [64] which is robust to outliers, i.e., to impulse noise. The parameter $\eta = 1$ enforces the differentiability of (33) with respect to u. Optimization of the functional is carried out using the Euler-Lagrange equations subject to Neumann boundary conditions:

$$\frac{\delta F_\varepsilon}{\delta v} = 2\beta v |\nabla u|^2 + \alpha \left(\frac{v - 1}{2\varepsilon} \right) - 2\varepsilon\alpha \Delta v = 0, \tag{34}$$

$$\frac{\delta F_\varepsilon}{\delta u} = \frac{(h * u - g)}{\sqrt{(h * u - g)^2 + \eta}} * h(-x, -y) - 2\beta \nabla \cdot \left(v^2 \nabla u \right) = 0. \tag{35}$$

The alternate minimization technique can be applied here as well since the functional (33) is convex, bounded from below and coercive with respect to either function u or v if the other one is fixed. Equation (34) is obviously linear with respect to v. In contrast, (35) is a nonlinear integrodifferential equation. Linearization of this equation is carried out using the fixed-point iteration scheme as in [33, 80]. In this method, additional iteration index l serves as intermediate stage calculating u^{n+1}. $u = u^l$ is set in the denominator, and $u = u^{l+1}$ elsewhere, where l is the current iteration number. Equation (35) can thus be rewritten as

$$H\left(v, u^l\right) u^{l+1} = G\left(u^l\right), l = 0, 1, \ldots. \tag{36}$$

where $\mathcal{H}$ is the linear integrodifferential operator

$$H\left(v, u^l\right) u^{l+1} = \frac{h * u^{l+1}}{\sqrt{\left(h * u^l - g\right)^2 + \eta}} * h(-x, -y) - 2\beta \nabla \cdot \left(v^2 \nabla u^{l+1} \right)$$

and

$$G\left(u^l\right) = \frac{g}{\sqrt{\left(h * u^l - g\right)^2 + \eta}} * h\left(-x, -y\right). \tag{37}$$

Note that (36) is now a *linear* integrodifferential equation in u^{l+1}.

The discretization of Eqs. (34) and (36) yields two systems of linear algebraic equations. These systems are solved in alternation, leading to the following iterative algorithm [12]:

Initialization: $u^0 = g$, $v^0 = 1$.

1. Solve for v^{n+1}

$$\left(2\beta \left|\nabla u^n\right|^2 + \frac{\alpha}{2\varepsilon} - 2\alpha\varepsilon\,\Delta\right) v^{n+1} = \frac{\alpha}{2\varepsilon}. \tag{38}$$

2. Set $u^{n+1,0} = u^n$ and solve for u^{n+1} (iterating on l)

$$H\left(v^{n+1}, u^{n+1,l}\right) u^{n+1,l+1} = G\left(u^{n+1,l}\right). \tag{39}$$

3. If $\left(\left\|u^{n+1} - u^n\right\|_{L_2} < \varepsilon_1 \left\|u^n\right\|_{L_2}\right)$, stop.

The convergence of the algorithm was proved in [14]. Figure 15 demonstrates the performance of the algorithm. The top row shows the blurred images with increasing salt-and-pepper noise level. The outcome of the restoration algorithm is shown in the bottom row.

A variant of the Mumford-Shah functional in its Γ-convergence approximation was suggested by Shah [74]. In this version, the L^2 norm of $|\nabla u|$ in (25) was replaced by its L^1 norm in the first term of G_ε

$$J_{\mathrm{MSTV}}(u, v) = \beta \int_\Omega v^2 |\nabla u|\, dx + \alpha \int_\Omega \left(\varepsilon |\nabla v|^2 + \frac{(v-1)^2}{4\varepsilon}\right) dx.$$

Alicandro et al. [2] proved the Γ-convergence of this functional to

$$J_{\mathrm{MSTV}}(u) = \beta \int_{\Omega/K} |\nabla u|\, dx + \alpha \int_K \frac{\left|u^+ - u^-\right|}{1 + \left|u^+ - u^-\right|} dH^1 + \left|D^c u\right|(\Omega),$$

where u^+ and u^- denote the image values on two sides of the edge set K, $\mathcal{H}^1$ is the one-dimensional Hausdorff measure, and $D^c u$ is the Cantor part of the measure-valued derivative Du. The Mumford-Shah and Shah regularizers are compared in Fig. 16. The blurred and noisy images are shown in the left column. The results of the restoration using the Mumford-Shah stabilizer (MS) are presented in the middle column, and the images recovered using the Shah regularizer (MSTV) are shown in the right column.

Fig. 15 *Top row*: the *Lena* image blurred with a pillbox kernel of radius 3 and contaminated by salt-and-pepper noise. The noise density is (*left* to *right*) 0.01, 0.1, and 0.3. *Bottom row*: the corresponding recovered images

The recovery using both methods is satisfactory, but it can be clearly seen that while the Mumford-Shah restoration performs better in the high-frequency image content (see the shades, for instance), the Shah restoration attracts the image toward the piecewise constant or cartoon limit which yields images much closer to the "ideal." This can be explained by the fact that the Shah regularizer is more robust to image gradients and hence eliminates high-frequency contributions.

The special case of pure impulse denoising (no blur) is demonstrated in Fig. 17. The image on the left shows the outcome of the algorithm of [65] with L^1 norm for both the fidelity and regularization, while the recovery using the L^1 fidelity and MS regularizer is shown on the right. It can be observed that the better robustness of the MS regularizer leads to better performance in the presence of salt-and-pepper noise.

16 Color Image Restoration

The restoration problem is now extended to vector-valued images [9]. In the case of color images, the image intensity is defined as u: $\Omega \to [0, 1]^3$. Here g^ν denotes the observed image at channel $\nu \in \{r, g, b\}$ such that $g^\nu = h * u^\nu + n^\nu$. The underlying assumption here is that the blur kernel h is common to all of the channels. If the noise is randomly located in a random color channel, the fidelity term can be

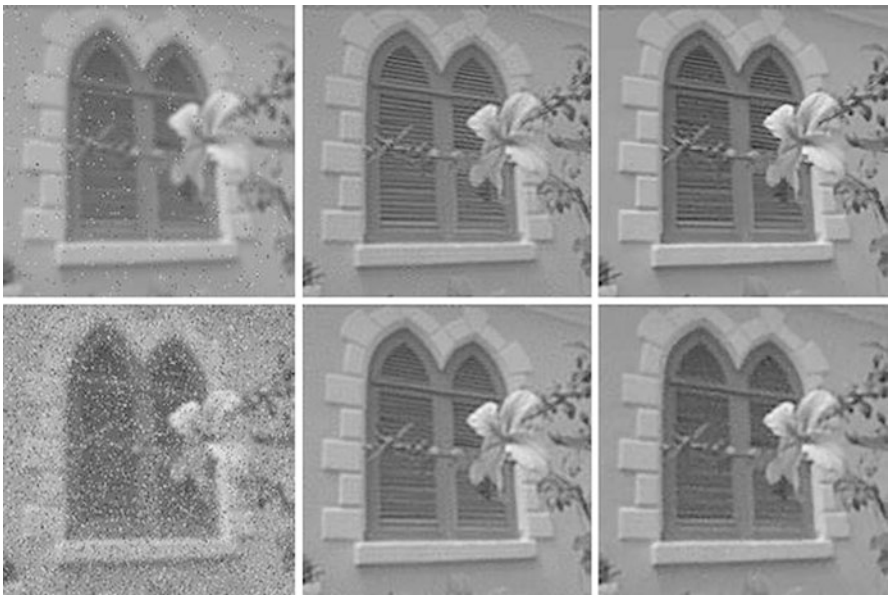


Fig. 16 *Left column*: the *window* image blurred with a pillbox kernel of radius 3 and contaminated by salt-and-pepper noise. The noise density is (*top* to *bottom*) 0.01 and 0.1. *Middle column*: the corresponding recovered images with Mumford-Shah (MS) regularization. *Right column*: the corresponding recovered images with Shah (MSTV) regularization

Fig. 17 Pure impulse denoising. *Left column*: restoration using the L^1 regularization [65]. *Right column*: restoration using the MS regularizer

modeled as

$$\Phi_{L^2} = \int_\Omega \sum\nolimits_\nu (h * u^\nu - g^\nu)^2 dx$$

in the case of Gaussian noise, and

$$\Phi_{L^1} = \int_\Omega \sqrt{(h * u^\nu - g^\nu)^2 + \eta}\, dx, \qquad \eta \ll 1, \tag{40}$$

in the case of impulsive noise. The TV regularization can be generalized to

$$\mathcal{J}_{TV}(u) = \int_{\Omega} \|\nabla u\| \, dx, \tag{41}$$

where

$$\|\nabla u\| = \sqrt{\sum\nolimits_{v\in\{r,g,b\}} |\nabla u^v|^2 + \mu}, \quad \mu \ll 1. \tag{42}$$

The color MS regularizer thus takes the form

$$\mathcal{J}_{\mathrm{MS}}(u, v) = \beta \int_{\Omega} v^2 \|\nabla u\|^2 \, dx + \alpha \int_{\Omega} \left(\varepsilon |\nabla v|^2 + \frac{(v-1)^2}{4\varepsilon} \right) dx. \tag{43}$$

Note that in this regularizer the edge map v is common for the three channels and provides the necessary coupling between colors. In the same fashion, the color MSTV regularizer is given by

$$\mathcal{J}_{\mathrm{MSTV}}(u, v) = \beta \int_{\Omega} v^2 \|\nabla u\| \, dx + \alpha \int_{\Omega} \left(\varepsilon |\nabla v|^2 + \frac{(v-1)^2}{4\varepsilon} \right) dx. \tag{44}$$

Once again, the optimization technique is alternate minimization with respect to u^v and v [9]. Figure 18 demonstrates the outcome of the different regularizers for an image blurred by Gaussian kernel and corrupted by both Gaussian and salt-and-pepper noise. The fidelity term in all cases was selected as Φ_L^1 (40).

The methods based on Mumford-Shah regularizer are superior to the TV stabilizers, where MSTV provides a result slightly closer to the "ideal" with little loss of details.

17 Space-Variant Restoration

The assumption of space-invariant blur kernel is sometimes inaccurate in real photographic images. For example, when multiple objects move at different velocities and in different directions in a scene, one gets space-variant motion blur. Likewise, when a camera lens is focused on one specific object, other objects nearer or farther away from the lens are not as sharp. In such situations, different blur kernels degrade different areas of the image. In some cases, it can be assumed that the blur kernel is a piecewise space-variant function. This means that every sub-domain in the image is blurred by a different kernel. In the full blind restoration, several operations have to be simultaneously applied: (1) segmentation of the subregions, (2) estimation of the blur kernels, and (3) recovery of the "ideal" image. Presented here is the simplest case where it is assumed that the subregions and blur kernels are known in advance. The segmentation procedure in a semi-blind restoration problem can be found in

Fig. 18 Recovery of the *Lena* image blurred by 7×7 out-of-focus kernel contaminated by mixture of Gaussian and salt-and-pepper noise

[15]. The non-blind space-variant restoration approach relies on the use of a global regularizer, which eliminates the requirement of dealing with region boundaries. As a result, the continuity of the gray levels in the recovered image is inherent. This method does not limit the number of subregions, their geometrical shape, and the kernel support size.

Let the open nonoverlapping subsets $w_i \subset \Omega$ denote regions that are blurred by kernels h_i, respectively. In addition, $\Omega/\cup\overline{w_i}$ denotes the background region blurred by the background kernel h_b, and $\overline{w_i}$ stands for the closure of w_i. The region boundaries are denoted by ∂w_i. The recovered image u is the minimizer of the objective functional

$$\mathcal{F}(u,v) = \frac{1}{2}\sum\nolimits_i \eta_i \int_{\omega_i} (h_i * u - g)^2 dx + \frac{\eta_b}{2}\int_{\Omega/(\cup\overline{\omega_i})} (h_b * u - g)^2 dx + \mathcal{J}_M. \tag{45}$$

where η_i and η_b are positive scalars and $\vartheta_{\mathrm{MS}}(u,v)$ is the Mumford-Shah regularizer (25). Following the formulation of Chan and Vese [31], the domains w_i can be

Fig. 19 Non-blind space-variant restoration. *Left column*: spatially variant motion blurred images. *Right column*: the corresponding recovered images using the suggested method

replaced by the Heaviside function $H(\phi_i)$, where

$$H(\phi_i) = \begin{cases} 1, \phi_i > 0, \\ 0, \phi_i \leq 0, \end{cases} \tag{46}$$

and $\phi_i : \Omega \to \mathbb{R}$ is a level set function such that

$$\partial\omega_i = \{x \in \Omega : \phi_i(x) = 0\}.$$

The functional then takes the form

$$\begin{aligned} \mathcal{F}(u, v) = \tfrac{1}{2} \textstyle\sum_i \eta_i \int_\Omega (h_i * u - g)^2 H(\phi_i) dx + \\ \tfrac{\eta_b}{2} \int_\Omega (h_b * u - g)^2 (1 - \textstyle\sum_i H(\phi_i)) dx + \mathcal{J}_{\mathrm{MS}}(u, v). \end{aligned} \tag{47}$$

Figure 19 demonstrates the performance of the suggested algorithm. The two images in the left column were synthetically blurred by different blur kernels within the marked shapes. The corresponding recovered images are shown in the right column. Special handling of the region boundaries was not necessary because the

MS regularizer was applied globally to the whole image, enforcing the piecewise-smoothness constraint. This means that the boundaries of the blurred regions were smoothed within the restoration process while edges were preserved.

18 Level Set Formulations for Joint Restoration and Segmentation

Presented here are other joint formulations for denoising, deblurring, and piecewise-constant segmentation introduced in [45] that can be seen as applications and modifications of the piecewise-constant Mumford-Shah model in level set formulation presented in Sect. 10. For related discussion, the reader is referred to [11–13, 47, 53]. A minimization approach is used, and the gradient descent method is considered. Let $g = h * u + n$ be a given blurred-noisy image, where h is a known blurring kernel (such as the Gaussian kernel) and n represents Gaussian additive noise of zero mean. We assume that the contours or jumps in the image u can be represented by the m distinct levels $\{-\infty = l_0 < l_1 < l_2 < \cdots < l_m < l_{m+1} = \infty\}$ of the same implicit (Lipschitz-continuous) function $\phi{:}\Omega \to \mathbb{R}$ partitioning Ω into $m+1$ disjoint open regions $R_j = \{x \in \Omega: l_{j-1} < \phi(x) < l_j\}$, $1 \le j \le m+1$. Thus, the denoised-deblurred image $u = c_1 H(\phi - l_m) + \Sigma_{j=2}^{m} c_j H(\phi - l_{m-j+1}) H(l_{m-j+2} - \phi) + c_{m+1} H(l_1 - \phi)$ is recovered by minimizing the following energy functional ($\nu_0 > 0$):

$$E(c_1, c_2, \ldots, c_{m+1}, \phi) = \int_\Omega |g - h * (c_1 H(\phi - l_m) + \sum_{j=2}^{m} c_j H(\phi - l_{m-j+} \\ + c_{m+1} H(l_1 - \phi))|^2 dx + \nu_0 \sum_{j=1}^{m} \int_\Omega |\nabla H(\phi - l_j)$$

In the binary case (one level $m = 1, l_1 = 0$), we assume the degradation model $g = h * c_1 H(\phi) + c_2(1 - H(\phi)) + n$, and we wish to recover $u = c_1 H(\phi) + c_2(1 - H(\phi))$ in Ω together with a segmentation of g. The modified binary segmentation model incorporating the blur becomes

$$\inf_{c_1, c_2, \phi} \{E(c_1, c_2, \phi) = \int_\Omega |g - h * (c_1 H(\phi) + c_2(1 - H(\phi)))|^2 dx \\ + \nu_0 \int_\Omega |\nabla H(\phi)| dx\}. \tag{48}$$

The Euler-Lagrange equations are computed minimizing this energy with respect to c_1, c_2, and ϕ. Using alternating minimization, keeping first ϕ fixed and minimizing the energy with respect to the unknown constants c_1 and c_2, the following linear system of equations are obtained:

$$c_1 \int_\Omega h_1^2 dx + c_2 \int_\Omega h_1 h_2 dx = \int g h_1 dx, \\ c_1 \int_\Omega h_1 h_2 dx + c_2 \int_\Omega h_2^2 dx = \int g h_2 dx$$

with the notations $h_1 = h * H(\phi)$ and $h_2 = h * (1 - H(\phi))$. Note that the linear system has a unique solution because the determinant of the coefficient matrix is not

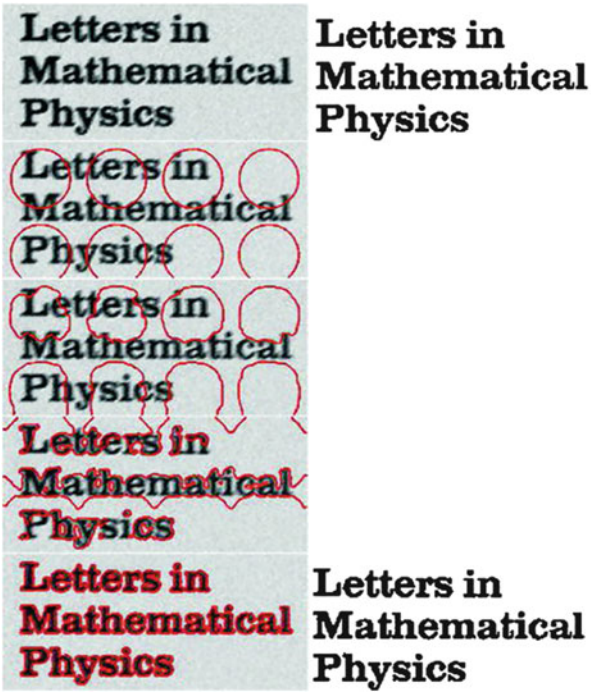

Letters in Mathematical Physics
Letters in Mathematical Physics
Letters in Mathematical Physics
Letters in Mathematical Physics
Letters in Mathematical Physics
Letters in Mathematical Physics
Letters in Mathematical Physics

Fig. 20 Joint segmentation, denoising, and deblurring using the binary level set model. *Top row*: (from *left* to *right*) degraded image g (blurred with motion blur kernel of length 10, oriented at an angle $\theta = 25°$ w.r.t. the horizon and contaminated by Gaussian noise with $\sigma_n = 10$), original image. Rows 2–5: initial curves, curve evolution using (48) at iterations 50, 100, and 300 with $\nu_0 = 5 \cdot 255^2$ and the restored image u(SNR = 28.1827). (c_1, c_2): original image $\approx$ (62.7525, 259.8939), restored u, (61.9194, 262.7795)

zero due to the Cauchy-Schwartz inequality $\int_\Omega h_1 h_2 \, dx^2 \leq \int_\Omega h_1^2 \, dx \int_\Omega h_2^2 \, dx$, where the equality holds if and only if $h_1 = h_2$ for a.e. $x \in \Omega$. But clearly, $h_1 = h * H(\phi))$ and $h_2 = h * (1 - H(\phi))$ are distinct; thus, we have strict inequality.

Keeping now the constants c_1 and c_2 fixed and minimizing the energy with respect to ϕ, the evolution equation is obtained by introducing an artificial time for the gradient descent in $\phi(t, x)$, $t > 0$, $x \in \Omega$

$$\frac{\partial \phi}{\partial t}(t, x) = \delta(\phi) \left[\left(\tilde{h} * g - c_1 \tilde{h} * (h * H(\phi)) - c_2 \tilde{h} * (h * (1 - H(\phi))) \right) (c_1 - c_2) + \nu_0 \operatorname{div} \left(\frac{\nabla \phi}{|\nabla \phi|} \right) \right],$$

where $\tilde{h}(x) = h(-x)$.

Figure 20 shows a numerical result for joint denoising, deblurring, and segmentation of a synthetic image, in a binary level set approach.

In the case of two distinct levels $l_1 < l_2$ of the level set function $\phi (m = 2)$, we wish to recover a piecewise-constant image of the form $u = c_1 H(\phi - l_2) + c_2 H(l_2 - \phi) H(\phi - l_1) + c_3 H(l_1 - \phi)$ and a segmentation of g, assuming the degradation model $g = h * c_1 H(\phi - l_2) + c_2 H(l_2 - \phi) H(\phi - l_1) + c_3 H(l_1 - \phi) + n$, by minimizing

$$\inf_{c_1, c_2, c_3, \phi} E(c_1, c_2, c_3, \phi) = \int_\Omega |g - h * (c_1 H(\phi - l_2) + c_2 H(l_2 - \phi) H(\phi - l_1) + c_3 H(l_1 - \phi))|^2 \, dx + \nu_0 \sum_{j=1}^{2} \int_\Omega |\nabla H(\phi - l_j)| \, dx. \tag{49}$$

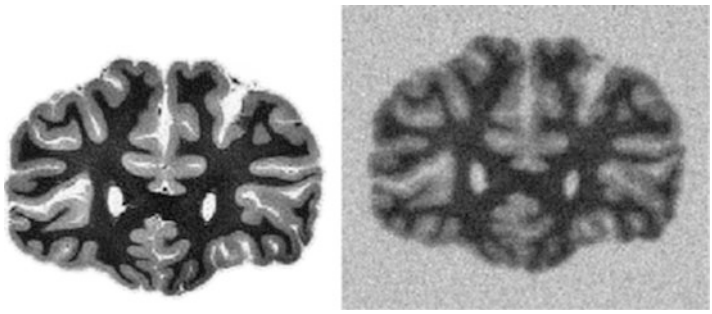


Fig. 21 Original image (*left*) and its noisy, blurry version (*right*) blurred with Gaussian kernel with $\sigma_b = 1$ and contaminated by Gaussian noise $\sigma_n = 20$

Similar to the previous binary model with blur, for fixed ϕ, the unknown constants are computed by solving the linear system of three equations:

$$\begin{aligned} c_1 \int h_1^2 dx + c_2 \int h_1 h_2 dx + c_3 \int h_1 h_3 dx &= \int g h_1 dx \\ c_1 \int h_1 h_2 dx + c_2 \int h_2^2 dx + c_3 \int h_2 h_3 dx &= \int g h_2 dx \\ c_1 \int h_1 h_3 dx + c_2 \int h_2 h_3 dx + c_3 \int h_3^2 dx &= \int g h_3 dx \end{aligned}$$

where $h_1 = h * H(\phi - l_2)$, $h_2 = h * H(l_2 - \phi)H(\phi - l_1)$, and $h_3 = h * H(l_1 - \phi)$.

For fixed c_1, c_2, and c_3, by minimizing the functional E with respect to ϕ, the gradient descent is obtained for $\phi(t, x)$, $t > 0$, $x \in \Omega$:

$$\begin{aligned} \frac{\partial \phi}{\partial t}(t,x) = \tilde{h} * (g - h * (&c_1 H(\phi - l_2) + c_2 H(l_2 - \phi) H(\phi - l_2) \\ &+ c_3 H(l_1 - \phi)) (c_1 \delta(\phi - l_2) \\ &+ c_2 H(l_2 - \phi) \delta(\phi - l_1) - c_2 H(\phi - l_1) \delta(l_2 - \phi) - c_3 \delta(l_1 - \phi))) \\ &+ \nu_0 \operatorname{div}\left(\frac{\nabla \phi}{|\nabla \phi|}\right) (\delta(\phi - l_1) + \delta(\phi - l_2)). \end{aligned} \tag{50}$$

Figures 21 and 22 show a numerical result for joint denoising, deblurring, and segmentation of the brain image in a multilayer level set approach.

19 Image Restoration by Nonlocal Mumford-Shah Regularizers

The traditional regularization terms discussed in the previous sections (depending on the image gradient) are based on local image operators, which denoise and preserve edges very well, but may induce loss of fine structures like texture during the restoration process. Recently, Buades et al. [22] introduced the nonlocal means filter, which produces excellent denoising results. Gilboa and Osher [42, 43] formulated the variational framework of NL means by proposing nonlocal regularizing functionals and the nonlocal operators such as the nonlocal gradient and

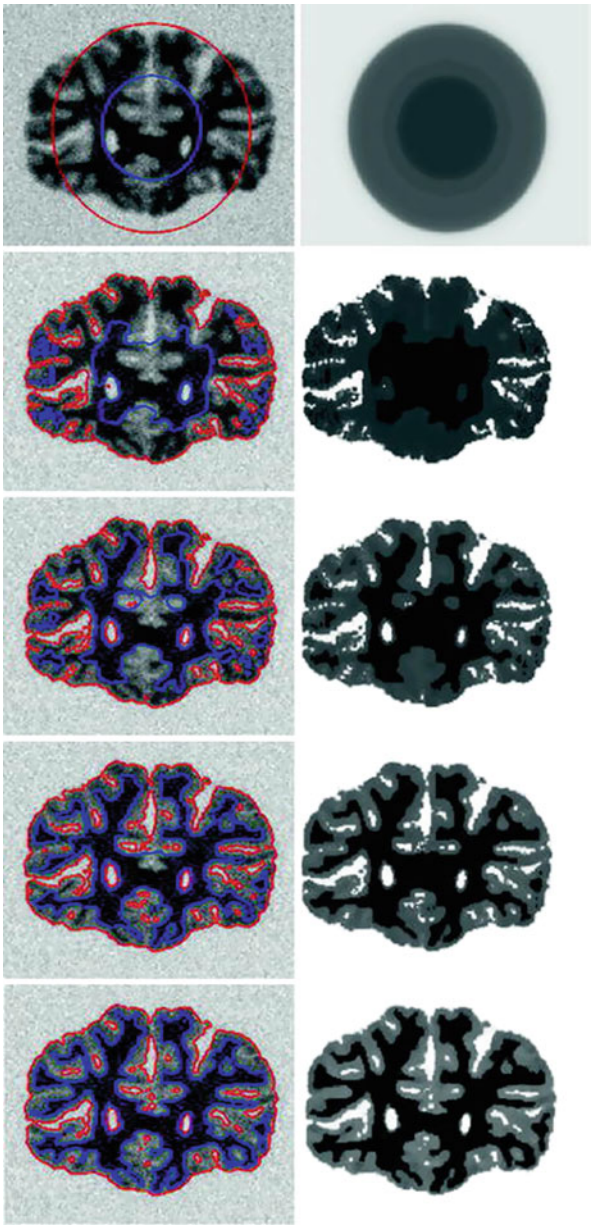


Fig. 22 Curve evolution and restored u using (51), $v_0 = 0.02 \cdot 255^2$, (c_1, c_2, c_3): original image $\approx$ (12.7501, 125.3610, 255.6453), restored $u \approx$ (22.4797, 136.9884, 255.0074)

divergence. Following Jung et al. [46], presented here are nonlocal versions of the Mumford-Shah-Ambrosio-Tortorelli regularizing functionals, called NL/MSH^1 and NL/MSTV, by applying the nonlocal operators proposed by Gilboa-Osher to MSH^1 and MSTV, respectively, for image restoration in the presence of blur and Gaussian or impulse noise. In addition, for the impulse noise model, use of a preprocessed

image is proposed to compute the weights w (the weights w defined in the NL means filter are more appropriate for the additive Gaussian noise case).

First recall the Ambrosio-Tortorelli regularizer,

$$\Psi_{\varepsilon}^{\mathrm{MSH}^1}(u,v)=\beta\int_{\Omega}v^2\left|\nabla u\right|^2dx+\alpha\int_{\Omega}\left(\varepsilon\left|\nabla v\right|^2+\frac{(v-1)^2}{4\varepsilon}\right)dx, \tag{51}$$

where $0\leq v(x)\leq 1$ represents the edges: $v(x)\approx 0$ if $x\in K$ and $v(x)\approx 1$ otherwise, ε is a small positive constant, and α, β are positive weights.

Shah [74] suggested a modified version of the approximation (51) to the MS functional by replacing the norm square of $|\nabla u|$ by the norm in the first term:

$$\Psi_{\varepsilon}^{\mathrm{MSTV}}(u,v)=\beta\int_{\Omega}v^2\left|\nabla u\right|dx+\alpha\int_{\Omega}\left(\varepsilon\left|\nabla v\right|^2+\frac{(v-1)^2}{4\varepsilon}\right)dx.$$

This functional Γ converges to the other ψ^{MSTV} functional [2]:

$$\Psi^{\mathrm{MSTV}}(u)=\beta\int_{\Omega/K}\left|\nabla u\right|dx+\alpha\int_{K}\frac{\left|u^+-u^-\right|}{1+\left|u^+-u^-\right|}dH^1+\left|D_cu\right|(\Omega),$$

where u^+ and u^- denote the image values on two sides of the jump set $K=J_u$ of u and D_cu is the Cantor part of the measure-valued derivative Du.

Nonlocal methods in image processing have been explored in many papers because they are well adapted to texture denoising, while the standard denoising models working with local image information seem to consider texture as noise, which results in losing texture. Nonlocal methods are generalized from the neighborhood filters and patch-based methods. The idea of neighborhood filter is to restore a pixel by averaging the values of neighboring pixels with a similar gray level value.

Buades et al. [22] generalized this idea by applying the patch-based methods, proposing a famous neighborhood filter called nonlocal means (or NL means):

$$\begin{aligned}NLu(x)&=\tfrac{1}{C(x)}\textstyle\int_{\Omega}e^{-\frac{d_a(u(x),u(y))}{h^2}}u(y)\,dy\\ d_a(u(x),u(y))&=\textstyle\int_{\mathbb{R}^2}G_a(t)\left|u(x+t)-u(y+t)\right|^2dt\end{aligned}$$

where d_a is the patch distance, G_a is the Gaussian kernel with standard deviation a determining the patch size, $C(x)=\int_{\Omega}e^{-\frac{d_a(u(x),u(y))}{h^2}}dy$ is the normalization factor, and h is the filtering parameter which corresponds to the noise level; it is usually set to be the standard deviation of the noise. The NL means not only compares the gray level at a single point but the geometrical configuration in a whole neighborhood (patch). Thus, to denoise a pixel, it is better to average the nearby pixels with similar structures rather than just with similar intensities.

In practice, the search window $\Omega_w = \{y \in \Omega: |y - x| \leq r\}$ is used instead of Ω (semi-local) and the weight function at $(x, y) \in \Omega \times \Omega$ depending on a function $u{:}\Omega \to \mathbb{R}$

$$w(x, y) = \exp\left(-\frac{d_a(u(x), u(y))}{h^2}\right).$$

The weight function $w(x, y)$ gives the similarity of image features between two pixels x and y, which is normally computed based on the blurry-noisy image g.

Based on the gradient and divergence definitions on graphs in the context of machine learning, Gilboa and Osher [43] derived the nonlocal operators. Let $u{:}\Omega \to \mathbb{R}$ be a function, and $w{:}\Omega \times \Omega \to \mathbb{R}$ is a weight function assumed to be nonnegative and symmetric. The nonlocal gradient $\nabla_w u{:}\Omega \times \Omega \to \mathbb{R}$ is defined as the vector $(\nabla_w u)(x, y) := (u(y) - u(x))\sqrt{w(x, y)}$. Hence, the norm of the nonlocal gradient of u at $x \in \Omega$ is defined as

$$|\nabla_w u|(x) = \sqrt{\int_\Omega (u(y) - u(x))^2 w(x, y)\, dy}.$$

The nonlocal divergence $\mathrm{div}_w \vec{v}{:}\Omega \to \mathbb{R}$ of the vector $\vec{v}{:}\Omega \times \Omega \to \mathbb{R}$ is defined as the adjoint of the nonlocal gradient

$$\left(\mathrm{div}_w \vec{v}\right)(x) := \int_\Omega (v(x, y) - v(y, x))\sqrt{w(x, y)}dy.$$

Based on these nonlocal operators, they introduced nonlocal regularizing functionals of the general form

$$\Psi(u) = \int_\Omega \phi\left(|\nabla_w u|^2\right) dx,$$

where $\phi(s)$ is a positive function, convex in $\sqrt{s}$ with $\phi(0) = 0$. Inspired by these ideas, nonlocal versions of Ambrosio-Tortorelli and Shah approximations to the MS regularizer for image denoising-deblurring are presented. This is also continuation of work by Bar et al. [11–13], as presented in the first part of this section.

Proposed are the following nonlocal approximated Mumford-Shah and Ambrosio-Tortorelli regularizing functionals (NL/MS) by applying the nonlocal operators to the approximations of the MS regularizer,

$$\Psi^{\mathrm{NL/MS}}(u, v) = \beta \int_\Omega v^2 \phi\left(|\nabla_w u|^2\right) dx + \alpha \int_\Omega \left(\varepsilon |\nabla v|^2 + \frac{(v-1)^2}{4\varepsilon}\right) dx,$$

where $\phi(s) = s$ and $\phi(s) = \sqrt{s}$ correspond to the nonlocal version of MSH[1] and MSTV regularizers, called here NL/MSH[1] and NL/MSTV, respectively:

$$\Psi^{\mathrm{NL}/\mathrm{MSH}^1}(u,v) = \beta \int_\Omega v^2 \left|\nabla_w u\right|^2 dx + \alpha \int_\Omega \left(\varepsilon \left|\nabla v\right|^2 + \frac{(v-1)^2}{4\varepsilon}\right) dx$$

$$\Psi^{\mathrm{NL}/\mathrm{MSTV}}(u,v) = \beta \int_\Omega v^2 \left|\nabla_w u\right| dx + \alpha \int_\Omega \left(\varepsilon \left|\nabla v\right|^2 + \frac{(v-1)^2}{4\varepsilon}\right) dx.$$

In addition, these nonlocal regularizers are used to deblur images in the presence of Gaussian or impulse noise. Thus, by incorporating the proper fidelity term depending on the noise model, two types of total energies as
Gaussian noise model:

$$E^G(u,v) = \int_\Omega (g - h * u)^2 dx + \Psi^{\mathrm{NL}/\mathrm{MS}}(u,v),$$

Impulse noise model:

$$E^{\mathrm{Im}}(u,v) = \int_\Omega \left|g - h * u\right| dx + \Psi^{\mathrm{NL}/\mathrm{MS}}(u,v).$$

Minimizing these functionals in u and v, the Euler-Lagrange equations:
Gaussian noise model:

$$\begin{aligned}
\frac{\partial E^G}{\partial v} &= 2\beta v \phi\left(\left|\nabla_w u\right|^2\right) - 2\varepsilon\alpha\, \Delta v + \alpha\left(\tfrac{v-1}{2\varepsilon}\right) = 0,\\
\frac{\partial E^G}{\partial u} &= h^* * (h * u - g) + L^{\mathrm{NL}/\mathrm{MS}} u = 0
\end{aligned}$$

Impulse noise model:

$$\begin{aligned}
\frac{\partial E^{\mathrm{Im}}}{\partial v} &= 2\beta v \phi\left(\left|\nabla_w u\right|^2\right) - 2\varepsilon\alpha\, \Delta v + \alpha\left(\tfrac{v-1}{2\varepsilon}\right) = 0,\\
\frac{\partial E^{\mathrm{Im}}}{\partial u} &= h^* * \operatorname{sign}(h * u - g) + L^{\mathrm{NL}/\mathrm{MS}} u = 0,
\end{aligned}$$

where $h^*(x) = h(-x)$ and

$$\begin{aligned}
L^{\mathrm{NL}/\mathrm{MS}} u = -2 \int_\Omega (u(y) - u(x))\, w(x,y) \\
\left[\left(v^2(y)\, \phi'\left(\left|\nabla_w(u)\right|^2(y)\right)\right.\right. \\
\left.\left. + v^2(x)\, \phi'\left(\left|\nabla_w(u)\right|^2(x)\right)\right] dy
\end{aligned}$$

More specifically, the NL/MSH^1 and NL/MSTV regularizers give

$$\begin{aligned} L^{\mathrm{NL}/\mathrm{MSH}^1} u &= -2\nabla_w \cdot \left(v^2(x)\, \nabla_w u(x)\right) \\ &= -2 \int_\Omega (u(y) - u(x)) w(x, y) \\ &\quad \left[v^2(y) + v^2(x)\right] dy, \end{aligned}$$

$$\begin{aligned} L^{\mathrm{NL/MSTV}} u &= -\nabla_w \cdot \left(v^2(x) \frac{\nabla_w u(x)}{|\nabla_w u(x)|}\right) \\ &= -\int_\Omega (u(y) - u(x))\, w(x, y) \left[\frac{v^2(y)}{|\nabla_w u|(y)} + \frac{v^2(x)}{|\nabla_w u|(x)}\right] dy \end{aligned}$$

The energy functionals $E^G(u, v)$ and $E^{\mathrm{Im}}(u, v)$ are convex in each variable and bounded from below. Therefore, to solve two Euler-Lagrange equations simultaneously, the alternate minimization (AM) approach is applied: in each step of the iterative procedure, we minimize with respect to one function while keeping the other one fixed. Due to its simplicity, the explicit scheme for u based on the gradient descent method and the Gauss-Seidel scheme for v is used. Note that since both energy functionals are not convex in the joint variable, only a local minimizer may be computed. However, this is not a drawback in practice, since the initial guess for u in our algorithm is the data g.

Furthermore, to extend the nonlocal methods to the impulse noise case, a preprocessing step is needed for the weight function $w(x, y)$ since it is not possible to directly use the data g to compute w. In other words, in the presence of impulse noise, the noisy pixels tend to have larger weights than the other neighboring points, so it is likely to keep the noise value at such pixel. Thus, a simple algorithm is proposed to obtain first a preprocessed image f, which removes the impulse noise (outliers) as well as preserves the textures as much as possible. Basically, the median filter is used, well known for removing impulse noise. However, if one step of the median filter is applied, then the output may be too smoothed out. In order to preserve the fine structures as well as to remove the noise properly, the idea of Bregman

Fig. 23 Original and noisy-blurry images (noisy-blurry image using the pillbox kernel of radius 2 and Gaussian noise with $\sigma_n = 5$)

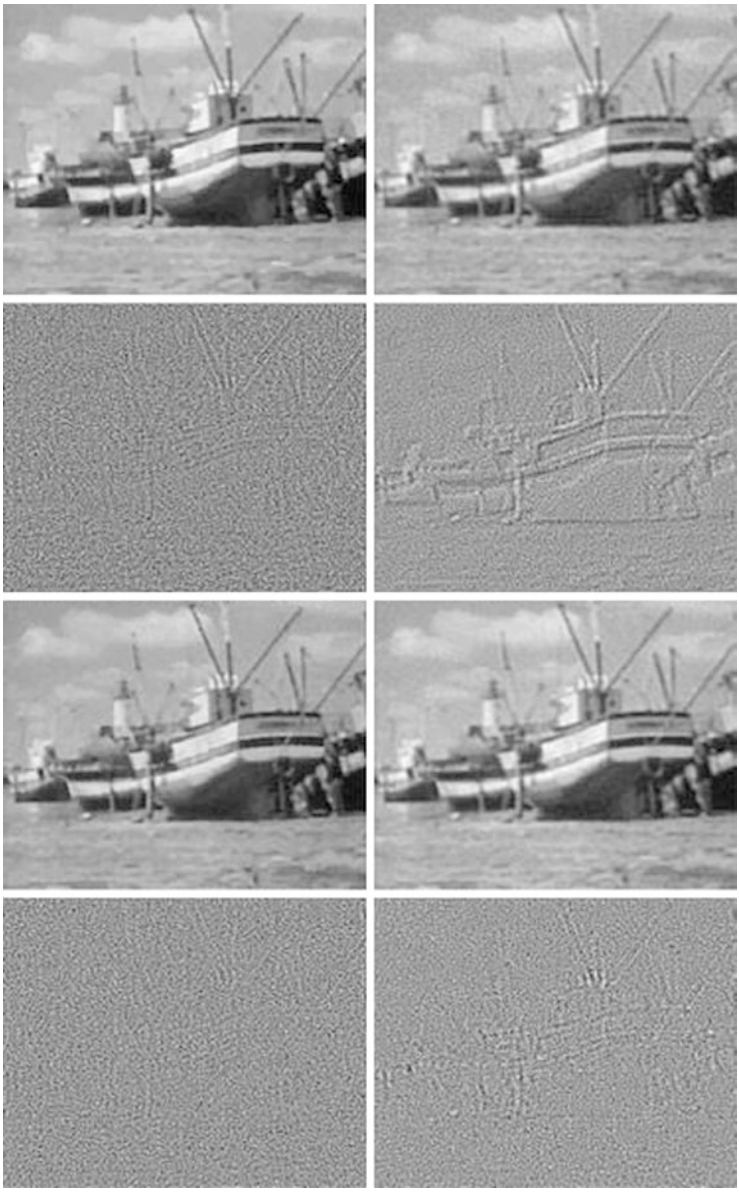


Fig. 24 Recovery of noisy-blurry image from Figs. 21–23. *Top row*: recovered image u using MSTV (SNR = 25.1968), MSH^1 (SNR = 23.1324). *Third row*: recovered image u using NL/MSTV (SNR = 26.4696), NL/MSH^1 (SNR = 24.7164). Second, *bottom rows*: corresponding residuals $g - h * u$. $\beta = 0.0045$ (MSTV), 0.001 (NL/MSTV), 0.06 (MSH^1), 0.006 (NL/MSH^1), $\alpha = 0.00000001$, $\varepsilon = 0.00002$

Fig. 25 Recovery of noisy-blurry image with Gaussian kernel with $\sigma = 1$ and salt-and-pepper noise with $d = 0.3$. *Top row*: original image, blurry image, noisy-blurry image. *Middle row*: recovered images using MSTV (SNR = 27.8336), MSH^1(SNR = 23.2052). *Bottom row*: recovered images using NL/MSTV (SNR = 29.3503), NL/MSH^1(SNR = 27.1477). *Parameters*: $\beta = 0.25$ (MSTV), 0.1 (NL/MSTV), $\alpha = 0.01$, $\varepsilon = 0.002$. *Parameters*: $\beta = 2$ (MSH^1), 0.55 (NL/MSH^1), $\alpha = 0.001$, $\varepsilon = 0.0001$

iteration is used [21,66], and the following algorithm is proposed to obtain a preprocessed image f that will be used only in the computation of the weight function:

- Initialize: $r_0 = 0$, $f_0 = 0$.
- do (iterate $n = 0, 1, 2, \ldots$)
- $f_{n+1} = \text{median}(g + r_n, [aa])$
- $r_{n+1} = r_n + g - h * f_{n+1}$
- while $||g - h_* f_n||_1 > ||g - h_* f_{n+1}||_1$
- Optional $f_m = \text{median}(f_m, [bb])$

where g is the given noisy-blurry data and *median* $(u, [aa])$ is the median filter of size $a \times a$ with input u; the optional step is needed in the case when the final f_m still has some salt-and-pepper-like noise. This algorithm is simple and requires a few iterations only, so it takes less than 1 s for a 256×256 size image. The preprocessed

Fig. 26 Edge map v using the MS regularizers in the recovery of the Lena image blurred with Gaussian blur kernel with $\sigma_b = 1$ and contaminated by salt-and-pepper noise with density $d = 0.3$. *Top*: (*left*) MSTV, (*right*) NL/MSTV. *Bottom*: (*left*) MSH^1, (*right*) NL/MSH^1

image f will be used only in the computation of the weights w while keeping g in the data-fidelity term; thus, artifacts are not introduced by the median filter.

Figures 23 and 24 show an experimental result for image restoration of a boat image degraded by the pillbox kernel blur of radius 2 and additive Gaussian noise. The nonlocal methods give better reconstruction.

Figures 25 and 26 show an experimental result for image restoration of a woman image degraded by Gaussian kernel blur and salt-and-pepper noise. Figure 26 shows the edge set v for the four results. The nonlocal methods give better reconstruction.

Figure 27 shows an experimental result for restoration of the Einstein image degraded by motion kernel blur and random-valued impulse noise. The nonlocal methods give better reconstruction.

20 Conclusion

This chapter is concluded by first summarizing its main results. The Mumford-Shah model for image segmentation has been presented, together with its main properties. Several approximations to the Mumford and Shah energy have been discussed, with an emphasis on phase-field approximations and level set approximations. Several numerical results for image segmentation by these methods have been presented. In the last section of the chapter, several restoration problems were addressed in a variational framework. The fidelity term was formulated according to the noise

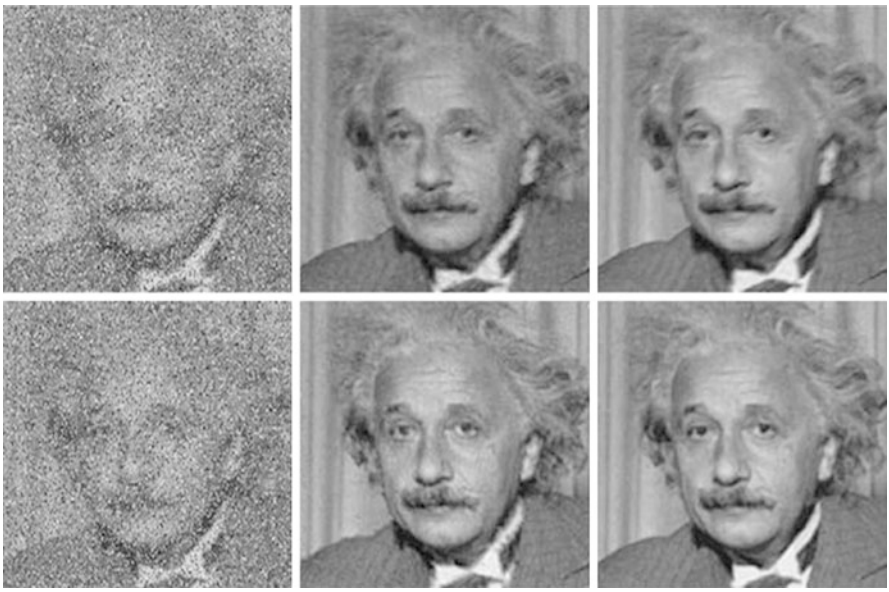


Fig. 27 Comparison between MSH^1 and NL/MSH^1 with the image blurred and contaminated by high density ($d = 0.4$) of random-valued impulse noise. *Top*: noisy-blurry image blurred with the motion blur in recovered images using MSH^1 (*left*, SNR = 17.9608) and NL/MSH^1 (*right*, SNR = 20.7563). *Bottom*: noisy blurry image blurred with the Gaussian blur in recovered images using MSH^1(*left*, SNR = 16.6960) and NL/MSH^1 (*right*, SNR = 24.2500). *Top*: $\beta = 1.5$ (MSH^1), 0.5 (NL/MSH^1), $\alpha = 0.0001$, $\varepsilon = 0.002$. *Bottom*: $\beta = 2.5$ (MSH^1), 0.65 (NL/MSH^1), $\alpha = 0.000001$, $\varepsilon = 0.002$

model (Gaussian, impulse, multichannel impulse). First, the a priori piecewise-smooth image model was mathematically integrated into the functional as an approximation of the Mumford-Shah segmentation elements by the Γ-convergence formulation. Comparative experimental results show the superiority of this regularizer with respect to modern state-of-the-art restoration techniques. Also, the piecewise-constant level set formulations of the Mumford-Shah energy have been applied to image restoration (related to relevant work by Kim et al. [47]), joint with segmentation. Finally, in the last section, the Ambrosio-Tortorelli approximations and Bar et al. restoration models have been extended to nonlocal regularizers, inspired by the work of Gilboa et al. These models produce much improved restoration results for images with texture and fine details.

21 Recommended Reading

Many more topics on the Mumford-Shah model and its applications have been explored in image processing, computer vision, and more generally inverse problems. This chapter contains only a small sample of results and methods. As

mentioned before, detailed monographs on the Mumford-Shah problem and related theoretical and application topics by Blake and Zisserman [16], by Morel and Solimini [60], by Chambolle [26], by Ambrosio et al. [4], by David [38], and by Braides [19] are recommended. Also, the monographs by Aubert and Kornprobst [8] and by Chan and Shen [28] contain chapters presenting the Mumford and Shah problem and its main properties.

The authors would like to mention the work by Cohen et al. [36, 37] on using curve evolution approach and the Mumford-Shah functional for detecting the boundary of a lake. The work by Aubert et al. [7] also proposes an interesting approximation of the Mumford-Shah energy by a family of discrete edge-preserving functionals, with Γ-convergence result.

Cross-References

- ▸ Duality and Convex Programming
- ▸ Energy Minimization Methods
- ▸ Level Set Methods for Structural Inversion and Image Reconstruction
- ▸ Mumford and Shah Model and its Applications to Image Segmentation and Image Restoration
- ▸ Total Variation in Imaging

References

1. Adams, R.A.: Sobolev Spaces. Academic, New York (1975)
2. Alicandro, R., Braides, A., Shah, J.: Free-discontinuity problems via functionals involving the L1-norm of the gradient and their approximation. Interfaces Free Bound **1**, 17–37 (1999)
3. Ambrosio, L.: A compactness theorem for a special class of functions of bounded variation. Boll. Un. Mat. Ital. **3**(B), 857–881 (1989)
4. Ambrosio, L., Fusco, N., Pallara, D.: Functions of Bounded Variation and Free Discontinuity Problems. Oxford University Press, New York (2000)
5. Ambrosio, L., Tortorelli, V.M.: Approximation of functionals depending on jumps by elliptic functionals via Γ-convergence. Commun. Pure Appl. Math. **43**(8), 999–1036 (1990)
6. Ambrosio, L., Tortorelli, V.M.: On the approximation of free discontinuity problems. Boll. Un. Mat. Ital. **B7**(6), 105–123 (1992)
7. Aubert, G., Blanc-Féraud, L., March, R.: An approximation of the Mumford-Shah energy by a family of discrete edge-preserving functionals. Nonlinear Anal. **64**(9), 1908–1930 (2006)
8. Aubert, G., Kornprobst, P.: Mathematical Problems in Image Processing. Springer, New York (2006)
9. Bar, L., Brook, A., Sochen, N., Kiryati, N.: Deblurring of color images corrupted by impulsive noise. IEEE Trans. Image Process. **16**(4), 1101–1111 (2007)
10. Bar, L., Sochen, N., Kiryati, N.: Variational pairing of image segmentation and blind restoration. In: Proceedings of 8th European Conference on Computer Vision, Prague. Volume 3022 of LNCS, pp. 166–177 (2004)
11. Bar, L., Sochen, N., Kiryati, N.: Image deblurring in the presence of salt-and-pepper noise. In: Proceedings of 5th International Conference on Scale Space and PDE Methods in Computer Vision, Hofgeismar. Volume 3459 of LNCS, pp. 107–118 (2005)
12. Bar, L., Sochen, N., Kiryati, N.: Image deblurring in the presence of impulsive noise. Int. J. Comput. Vis. **70**, 279–298 (2006)

13. Bar, L., Sochen, N., Kiryati, N.: Semi-blind image restoration via Mumford-Shah regularization. IEEE Trans. Image Process. **15**(2), 483–493 (2006)
14. Bar, L., Sochen, N., Kiryati, N.: Convergence of an iterative method for variational deconvolution and impulsive noise removal. SIAM J. Multiscale Model Simul. **6**, 983–994 (2007)
15. Bar, L., Sochen, N., Kiryati, N.: Restoration of images with piecewise space-variant blur. In: Proceedings of 1st International Conference on Scale Space and Variational Methods in Computer Vision, Ischia, pp. 533–544 (2007)
16. Blake, A., Zisserman, A.: Visual Reconstruction. MIT, Cambridge (1987)
17. Bourdin, B.: Image segmentation with a finite element method. M2AN Math. Model. Numer. Anal. **33**(2), 229–244 (1999)
18. Bourdin, B., Chambolle, A.: Implementation of an adaptive finite-element approximation of the Mumford-Shah functional. Numer. Math. **85**(4), 609–646 (2000)
19. Braides, A.: Approximation of Free-Discontinuity Problems. Volume 1694 of Lecture Notes in Mathematics. Springer, Berlin (1998)
20. Braides, A., Dal Maso, G.: Nonlocal approximation of the Mumford-Shah functional. Calc Var **5**, 293–322 (1997)
21. Bregman, L.M.: The relaxation method for finding common points of convex sets and its application to the solution of problems in convex programming. USSR Comput. Math. Phys. **7**, 200–217 (1967)
22. Buades, A., Coll, B., Morel, J.M.: A review of image denoising algorithms, with a new one. SIAM MMS **4**(2), 490–530 (2005)
23. Chambolle, A.: Un théorème de γ-convergence pour la segmentation des signaux. C R Acad Sci Paris Sér. I Math **314**(3), 191–196 (1992)
24. Chambolle, A.: Image segmentation by variational methods: Mumford and Shah functional, and the discrete approximation. SIAM J. Appl. Math. **55**, 827–863 (1995)
25. Chambolle, A.: Finite-differences discretizations of the Mumford-Shah functional. M2AN Math. Model. Numer. Anal. **33**(2), 261–288 (1999)
26. Chambolle, A.: Inverse problems in image processing and image segmentation: some mathematical and numerical aspects. In: Chidume, C.E. (ed.) Mathematical Problems in Image Processing. ICTP Lecture Notes Series, vol. 2. ICTP, Trieste (2000). http://publications.ictp.it/lns/vol2.html
27. Chambolle, A., Dal Maso, G.: Discrete approximation of the Mumford-Shah functional in dimension two. M2AN Math. Model. Numer. Anal. **33**(4), 651–672 (1999)
28. Chan, T.F., Shen, J.: Image Processing and Analysis. Variational, PDE, Wavelet, and Stochastic Methods. SIAM, Philadelphia (2005)
29. Chan, T., Vese, L.: An active contour model without edges. Lect. Notes Comput. Sci. **1682**, 141–151 (1999)
30. Chan, T., Vese, L.: An efficient variational multiphase motion for the Mumford-Shah segmentation model. In: 34th Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, vol. 1, pp. 490–494 (2000)
31. Chan, T., Vese, L.: Active contours without edges. IEEE Trans. Image Process. **10**, 266–277 (2001)
32. Chan, T., Vese, L.: A level set algorithm for minimizing the Mumford-Shah functional in image processing. In: IEEE/Computer Society Proceedings of the 1st IEEE Workshop on Variational and Level Set Methods in Computer Vision, Vancouver, pp. 161–168 (2001)
33. Chan, T.F., Wong, C.K.: Total variation blind deconvolution. IEEE Trans. Image Process. **7**, 370–375 (1998)
34. Chung, G., Vese, L.A.: Energy minimization based segmentation and denoising using a multilayer level set approach. Lect. Notes Comput. Sci. **3757**, 439–455 (2005)
35. Chung, G., Vese, L.A.: Image segmentation using a multilayer level-set approach. Comput. Vis. Sci. **12**(6), 267–285 (2009)
36. Cohen, L.D.: Avoiding local minima for deformable curves in image analysis. In: Le Méhauté, A., Rabut, C., Schumaker, L.L. (eds.) Curves and Surfaces with Applications in CAGD, pp. 77–84. Vanderbilt University Press, Nashville (1997)

37. Cohen, L., Bardinet, E., Ayache, N.: Surface reconstruction using active contour models. In: SPIE '93 Conference on Geometric Methods in Computer Vision, San Diego, July 1993
38. David, G.: Singular Sets of Minimizers for the Mumford-Shah Functional. Birkhäuser, Basel (2005)
39. Evans, L.C.: Partial Differential Equations. American Mathematical Society, Providence (1998)
40. Evans, L.C., Gariepy, R.F.: Measure Theory and Fine Properties of Functions. CRC, Boca Raton (1992)
41. Geman, S., Geman, D.: Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE TPAMI **6**, 721–741 (1984)
42. Gilboa, G., Osher, S.: Nonlocal linear image regularization and supervised segmentation. SIAM MMS **6**(2), 595–630 (2007)
43. Gilboa, G., Osher, S.: Nonlocal operators with applications to image processing. Multiscale Model. Simul. **7**(3), 1005–1028 (2008)
44. Huber, P.J.: Robust Statistics. Wiley, New York (1981)
45. Jung, M., Chung, G., Sundaramoorthi, G., Vese, L.A., Yuille, A.L.: Sobolev gradients and joint variational image segmentation, denoising and deblurring. In: IS&T/SPIE on Electronic Imaging. Volume 7246 of Computational Imaging VII, San Jose, pp. 72460I-1–72460I-13 (2009)
46. Jung, M., Vese, L.A.: Nonlocal variational image deblurring models in the presence of gaussian or impulse noise. In: International Conference on Scale Space and Variational Methods in Computer Vision (SSVM' 09), Voss. Volume 5567 of LNCS, pp. 402–413 (2009)
47. Kim, J., Tsai, A., Cetin, M., Willsky, A.S.: A curve evolution-based variational approach to simultaneous image restoration and segmentation. In: Proceedings of IEEE International Conference on Image Processing, Rochester, vol. 1, pp. 109–112 (2002)
48. Koepfler, G., Lopez, C., Morel, J.M.: A multiscale algorithm for image segmentation by variational methods. SIAM J. Numer. Anal. **31**(1), 282–299 (1994)
49. Kundur, D., Hatzinakos, D.: Blind image deconvolution. Signal Process. Mag. **13**, 43–64 (1996)
50. Kundur, D., Hatzinakos, D.: Blind image deconvolution revisited. Signal Process. Mag. **13**, 61–63 (1996)
51. Larsen, C.J.: A new proof of regularity for two-shaded image segmentations. Manuscr. Math. **96**, 247–262 (1998)
52. Leonardi, G.P., Tamanini, I.: On minimizing partitions with infinitely many components. Ann. Univ. Ferrara Sez. VII Sc. Mat. **XLIV**, 41–57 (1998)
53. Li, C., Kao, C.-Y., Gore, J.C., Ding, Z.: Implicit active contours driven by local binary fitting energy. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), CVPR'07, Minneapolis (2007)
54. Dal Maso, G.: An Introduction to Γ-Convergence. Progress in Nonlinear Differential Equations and Their Applications. Birkhäuser, Boston (1993)
55. Dal Maso, G., Morel, J.M., Solimini, S.: Variational approach in image processing – existence and approximation properties. C. R. Acad. Sci. Paris Sér. I Math. **308**(19), 549–554 (1989)
56. Dal Maso, G., Morel, J.M., Solimini, S.: A variational method in image segmentation – existence and approximation properties. Acta Math. **168**(1–2), 89–151 (1992)
57. Massari, U., Tamanini, I.: On the finiteness of optimal partitions. Ann. Univ. Ferrara Sez VII Sc. Mat. **XXXIX**, 167–185 (1993)
58. Modica, L.: The gradient theory of phase transitions and the minimal interface criterion. Arch. Ration. Mech. Anal. **98**, 123–142 (1987)
59. Modica, L., Mortola, S.: Un esempio di γ-convergenza. Boll. Un. Mat. Ital. **B5**(14), 285–299 (1977)
60. Morel, J.-M., Solimini, S.: Variational Methods in Image Segmentation. Birkhäuser, Boston (1995)
61. Mumford, D., Shah, J.: Boundary detection by minimizing functionals. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, pp. 22–26 (1985)

62. Mumford, D., Shah, J.: Boundary detection by minimizing functionals. In: Ullman, S., Richards, W. (eds.) Image Understanding, pp. 19–43. Springer, Berlin (1989)
63. Mumford, D., Shah, J.: Optimal approximations by piecewise smooth functions and associated variational problems. Commun. Pure Appl. Math. **42**, 577–685 (1989)
64. Nikolova, M.: Minimizers of cost-functions involving nonsmooth data-fidelity terms: application to the processing of outliers. SIAM J. Numer. Anal. **40**, 965–994 (2002)
65. Nikolova, M.: A variational approach to remove outliers and impulse noise. J. Math. Imaging Vis. **20**, 99–120 (2004)
66. Osher, S., Burger, M., Goldfarb, D., Xu, J., Yin, W.: An iterative regularization method for total variation based image restoration. SIAM MMS **4**, 460–489 (2005)
67. Osher, S.J., Fedkiw, R.P.: Level Set Methods and Dynamic Implicit Surfaces. Springer, New York (2002)
68. Osher, S., Sethian, J.A.: Fronts propagating with curvature-dependent speed: algorithms based on Hamilton-Jacobi formulation. J. Comput. Phys. **79**, 12–49 (1988)
69. Rudin, L., Osher, S.: Total variation based image restoration with free local constraints. In: Proceedings of IEEE International Conference on Image Processing, Austin, vol. 1, pp. 31–35 (1994)
70. Rudin, L.I., Osher, S., Fatemi, E.: Non linear total variation based noise removal algorithms. Physica D **60**, 259–268 (1992)
71. Samson, C., Blanc-Féraud, L., Aubert, G., Zerubia, J.: Multiphase evolution and variational image classification. Technical report 3662, INRIA Sophia Antipolis (1999)
72. Sethian, J.A.: Level Set Methods: Evolving Interfaces in Geometry, Fluid Mechanics, Computer Vision, and Materials Science. Cambridge Monograph on Applied and Computational Mathematics, Cambridge, United Kingdom, University Press, Cambridge (1996)
73. Sethian, J.A.: Level Set Methods and Fast Marching Methods. Evolving Interfaces in Computational Geometry, Fluid Mechanics, Computer Vision, and Materials Science. Cambridge University Press, Cambridge (1999)
74. Shah, J.: A common framework for curve evolution, segmentation and anisotropic diffusion. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, pp. 136–142 (1996)
75. Tamanini, I.: Optimal approximation by piecewise constant functions. In: Serapioni R., Tomarelli F. (eds.) Variational Methods for Discontinuous Structures: Applications to Image Segmentation, Continuum Mechanics, Homogenization, Villa Olmo, Como, 8–10 September 1994. Progress in Nonlinear Differential Equations and Their Applications, vol. 25, pp. 73–85. Birkhäuser, Basel (1996)
76. Tamanini, I., Congedo, G.: Optimal segmentation of unbounded functions. Rend. Sem. Mat. Univ. Padova **95**, 153–174 (1996)
77. Tikhonov, A.N., Arsenin, V.: Solutions of Ill-Posed Problems. Winston, Washington (1977)
78. Tsai, A., Yezzi, A., Willsky, A.: Curve evolution implementation of the Mumford-Shah functional for image segmentation, denoising, interpolation, and magnification. IEEE Trans. Image Process. **10**(8), 1169–1186 (2001)
79. Vese, L.A., Chan, T.F.: A multiphase level set framework for image segmentation using the Mumford and Shah model. Int. J. Comput. Vis. **50**(3), 271–293 (2002)
80. Vogel, C.R., Oman, M.E.: Fast, robust total variation-based reconstruction of noisy, blurred images. IEEE Trans. Image Process. **7**, 813–824 (1998)
81. Weisstein, E.W.: Minimal residual method. MathWorld-A Wolfram Web Resource. http://mathworld.wolfram.com/MinimalResidualMethod.html
82. You, Y., Kaveh, M.: A regularization approach to joint blur identification and image restoration. IEEE Trans. Image Process. **5**, 416–428 (1996)
83. Zhao, H.K., Chan, T., Merriman, B., Osher, S.: A variational level set approach to multiphase motion. J. Comput. Phys. **127**, 179–195 (1996)

Local Smoothing Neighborhood Filters

Jean-Michel Morel, Antoni Buades, and Tomeu Coll

Contents

1 Introduction 1600
2 Denoising 1607
Analysis of Neighborhood Filter as a Denoising Algorithm 1607
Neighborhood Filter Extension: The NL-Means Algorithm 1609
Extension to Movies 1614
3 Asymptotic 1614
PDE Models and Local Smoothing Filters 1614
Asymptotic Behavior of Neighborhood Filters (Dimension 1) 1619
The Two-Dimensional Case 1623
A Regression Correction of the Neighborhood Filter 1625
The Vector-Valued Case 1628
4 Variational and Linear Diffusion 1633
Linear Diffusion: Seed Growing 1634
Linear Diffusion: Histogram Concentration 1635
5 Conclusion 1639
Cross-References 1640
References 1640

J.-M. Morel (✉)
École Normale Supérieure de Cachan, Cachan, France
e-mail: moreljeanmichel@gmail.com

A. Buades
Universitat Illes Balears, Palma de Mallorca, Spain
e-mail: toni.buades@uib.es

T. Coll
Universitat de les Illes Balears, Palma-Illes Balears, Spain
e-mail: tomeu.coll@uib.es

© Springer Science+Business Media New York 2015

O. Scherzer (ed.), *Handbook of Mathematical Methods in Imaging*,
DOI 10.1007/978-1-4939-0790-8_26

Abstract

Denoising images can be achieved by a spatial averaging of nearby pixels. However, although this method removes noise, it creates blur. Hence, neighborhood filters are usually preferred. These filters perform an average of neighboring pixels, but only under the condition that their gray level is close enough to the one of the pixel in restoration. This very popular method unfortunately creates shocks and staircasing effects. It also excessivelly blurs texture and fine structures when noise dominates the signal.

In this chapter, we perform an asymptotic analysis of neighborhood filters as the size of the neighborhood shrinks to zero. We prove that these filters are asymptotically equivalent to the Perona-Malik equation, one of the first nonlinear PDEs proposed for image restoration. As a solution to the shock effect, we propose an extremely simple variant of the neighborhood filter using a linear regression instead of an average. By analyzing its subjacent PDE, we prove that this variant does not create shocks: it is actually related to the mean curvature motion.

We also present a generalization of neighborhood filters, the nonlocal means (NL-means) algorithm, addressing the preservation of structure in a digital image. The NL-means algorithm tries to take advantage of the high degree of redundancy of any natural image. By this, we simply mean that every small window in a natural image has many similar windows in the same image. Now in a very general sense inspired by the neighborhood filters, one can define as "neighborhood of a pixel" any set of pixels with a similar window around. All pixels in that neighborhood can be used for predicting its denoised value.

We finally analyze the recently introduced variational formulations of neighborhood filters and their application to segmentation and seed diffusion.

1 Introduction

The *neighborhood filter or sigma filter* is attributed to J.S. Lee [48] (in 1983) but goes back to L. Yaroslavsky and the Sovietic image-processing theory [75]. This filter is introduced in a denoising framework for the removal of additive white noise:

$$v(x) = u(x) + n(x),$$

where $\mathbf{x}$ indicates a pixel site, $v(\mathbf{x})$ is the noisy value, $u(\mathbf{x})$ is the "true" value at pixel $\mathbf{x}$, and $n(\mathbf{x})$ is the noise perturbation. When the noise values $n(\mathbf{x})$ and $n(\mathbf{y})$ at different pixels are assumed to be independent random variables and independent of the image value $u(x)$, one talks about "white noise." Generally, $n(\mathbf{x})$ is supposed to follow a Gaussian distribution of zero mean and standard deviation σ.

Lee and Yaroslavsky proposed to smooth the noisy image by averaging only those neighboring pixels that have a similar intensity. Averaging is the principle of most denoising methods. The variance law in probability theory ensures that if N

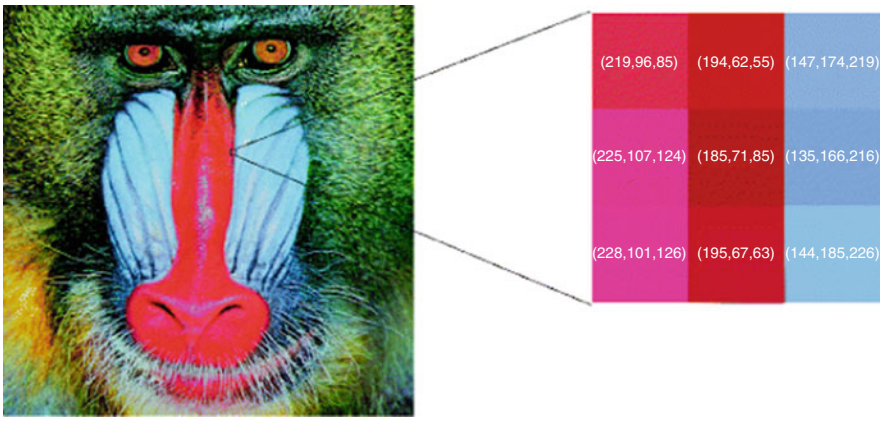


Fig. 1 The nine pixels in the baboon image on the *right* have been enlarged. They present a high *red-blue* contrast. In the *red pixels*, the first (*red*) component is stronger. In the *blue pixels*, the third component, *blue*, dominates

noise values are averaged, the noise standard deviation is divided by $\sqrt{N}$. Thus, one should, for example, find for each pixel nine other pixels in the image with the same color (up to the fluctuations due to noise) in order to reduce the noise by a factor 3. A first idea might be to chose the closest ones. Now, the closest pixels have not necessarily the same color as illustrated in Fig. 1. Look at the red pixel placed in the middle of Fig. 1. This pixel has five red neighbors and three blue ones. If the color of this pixel is replaced by the average of the colors of its neighbors, it turns blue. The same process would likewise redden the blue pixels of this figure. Thus, the red and blue border would be blurred. It is clear that in order to denoise the central red pixel, it is better to average the color of this pixel with the nearby red pixels and only them, excluding the blue ones. This is exactly the technique proposed by neighborhood filters.

The original sigma and neighborhood filter were proposed as an average of the spatially close pixels with a gray level difference lower than a certain threshold h. Thus, for a certain pixel $\mathbf{x}$, the denoised value is the average of pixels in the spatial and intensity neighborhood:

$$\{y \in \Omega \,|\, \|x - y\| < \rho \text{ and } |u(x) - u(y)| < h\}.$$

However, in order to make it coherent with further extensions and facilitate the mathematical development of this chapter, we will write the filter in a continuous framework under a weighted average form. We will denote the neighborhood or sigma filter by NF and define it for a pixel $\mathbf{x}$ as

$$\mathrm{NF}_{h,\rho}u(x) = \frac{1}{C(x)} \int_{B_\rho(x)} u(y) e^{-\frac{|u(y)-u(x)|^2}{h^2}} dy, \tag{1}$$

where $B_\rho(\mathbf{x})$ is a ball of center $\mathbf{x}$ and radius $\rho > 0, h > 0$ is the filtering parameter, and $C(x) = \int_{B_\rho(x)} e^{-\frac{|u(y)-u(x)|^2}{h^2}} dy$ is the normalization factor. The parameter h controls the degree of color similarity needed to be taken into account in the average. This value depends on the noise standard deviation σ, and it was set to 2.5σ in [48] and [75].

The Yaroslavsky and Lee's filter (1) is less known than more recent versions, namely, the *SUSAN filter* [68] and the *bilateral filter* [70]. Both algorithms, instead of considering a fixed spatial neighborhood $B_\rho(\mathbf{x})$, weigh the distance to the reference pixel $\mathbf{x}$:

$$\mathrm{SF}_{h,\rho}u(x) = \frac{1}{C(x)} \int_\Omega u(y) e^{-\frac{|y-x|^2}{\rho^2}} e^{-\frac{|u(y)-u(x)|^2}{h^2}} dy, \tag{2}$$

where $C(x) = \int_\Omega e^{-\frac{|y-x|^2}{\rho^2}} e^{-\frac{|u(y)-u(x)|^2}{h^2}} dy$ is the normalization factor and $\rho > 0$ is now a spatial filtering parameter. Even if the SUSAN algorithm was previously introduced, the whole literature refers to it as the bilateral filter. Therefore, we shall call this filter by the latter name in subsequent sections.

The only difference between the neighborhood filter and the bilateral or SUSAN filter is the way the spatial component is treated. While for the neighborhood filter all pixels within a certain spatial distance are treated uniformly, for the bilateral or SUSAN filter, pixels closer to the reference one are considered more important. We display in Fig. 2 a denoising experience where a Gaussian white noise of standard deviation 10 has been added to a non-noisy image. We display the denoised image by both the neighborhood and bilateral filters. We observe that both filters avoid the excessive blurring caused by a Gaussian convolution and preserve all contrasted edges in the image.

Fig. 2 From left to right: noise image, Gaussian convolution, neighborhood filter, and bilateral filter. The neighborhood and bilateral filters avoid the excessive blurring caused by a Gaussian convolution and preserve all contrasted edges in the image

The above denoising experience was applied to color images. In order to clarify how the neighborhood filters are implemented in this case, we remind that each pixel $\mathbf{x}$ is a triplet of values $u(\mathbf{x}) = (u_1(\mathbf{x}), u_2(\mathbf{x}), u_3(\mathbf{x}))$, denoting the red, green, and blue components. Then, the filter rewrites

$$\mathrm{NF}_{h,\rho}u_i(x) = \frac{1}{C(x)} \int_{B_\rho(x)} u_i(y) e^{-\frac{\|u(y)-u(x)\|^2}{h^2}} dy,$$

$||u(\mathbf{y}) - u(\mathbf{x})||^2$ being the average of the distances of the three channels:

$$\|u(y) - u(x)\|^2 = \frac{1}{3} \sum_{i=1}^{3} |u_i(y) - u_i(x)|.$$

The same definition applies for the SUSAN or bilateral filter by incorporating the spatial weighting term. The above definition naturally extends to multispectral images with an arbitrary number of channels. Bennett et al. [7] applied it to multispectral data with an infrared channel and Peng et al. [56] for general multispectral data.

The evaluation of the denoising performance of neighborhood filters and comparison with state-of-the-art algorithms are postponed to Sect. 2. In the same section, we present a natural extension of the neighborhood filter, the NL-means algorithm, proposed in [12]. This algorithm evaluates the similarity between two pixels $\mathbf{x}$ and $\mathbf{y}$ not only by the intensity or color difference of $\mathbf{x}$ and $\mathbf{y}$ but by the difference of intensities in a whole spatial neighborhood.

The bilateral filter was also proposed as a filtering algorithm with a filtering scale depending on both parameters h and ρ. Thus, taking several values for these parameters, we obtain different filtered images and corresponding residuals in a multi-scale framework. In Fig. 3, we display several applications of the bilateral filter for different values of the parameters h and ρ. We also display the differences between the original and filtered images in Fig. 4. For moderated values of h, this residual contains details and texture, but it does not contain contrasted edges. This contrasted information is removed by the bilateral filter only for large values of h. In that case, all pixels are judged as having a similar intensity level and the weight is set taking into account only the spatial component. It is well known that the residual by such an average is proportional to the Laplacian of the image. In Sect. 2, we will mathematically analyze the asymptotical expansion of the neighborhood residual image.

This detail removal of the bilateral while conserving very contrasted edges is the key in many image and video processing algorithms. Durand and Dorsey [28] use this property in the context of tone mapping whose goal is to compress the intensity values of a high-dynamic-range image. The authors isolate the details before compressing the range of the image. Filtered details and texture are added back at the final stage. Similar approaches for image editing are presented by Bae et al. [5], which transfer the visual look of an artist picture onto a casual photograph. Eisemann and Durand [32] and Petschnigg et al. [58] combine the filtered and

Fig. 3 Several applications of the bilateral filter for increasing values of parameters ρ and h. The parameter ρ increases from top to bottom taking values $\{2, 5, 10\}$ and h increases from left to right taking values $\{5, 10, 25, 100\}$

residual image of a flash and non-flash image of the same scene. These two last algorithms, in addition, compute the weight configuration in one image of the pair and average the intensity values of the other image. As we will see, this is a common feature with iterative versions of neighborhood filters. However, for these applications, both images of the pair must be correctly and precisely registered.

The iteration of the neighborhood filter was not originally considered by the pioneering works of Lee and Yaroslavsky. However, recent applications have shown its interest. The iteration of the filter as a local smoothing operator tends to piecewise

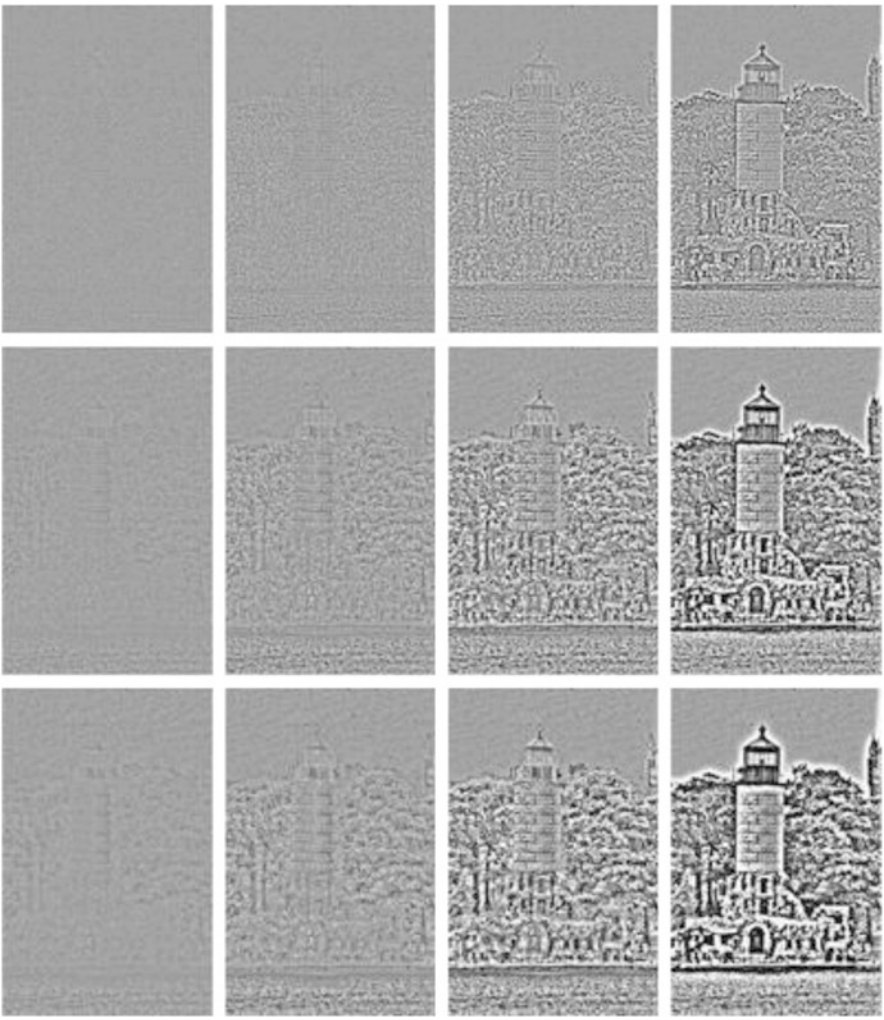


Fig. 4 Residual differences between original and filtered images in Fig. 3. For moderated values of h this residual contains details and texture but it doesn't contain contrasted edges. These contrasted information is removed by the bilateral filter only for large values of h

constant images by creating artificial discontinuities in regular zones. Barash et al. [6] showed that an iteration of the neighborhood filter was equivalent to a step of a certain numerical scheme of the classical Perona-Malik equation [57]. A complete proof of the equivalence between the neighborhood filter and the Perona-Malik equation was presented in [13] including a modification of the filter to avoid the creation of shocks inside regular parts of the image. Another theoretical explanation of the shock effect of the neighborhood filters can be found in Van de Weijer and van den Boomgaard [71] and Comaniciu [20]. Both papers show that the iteration of the

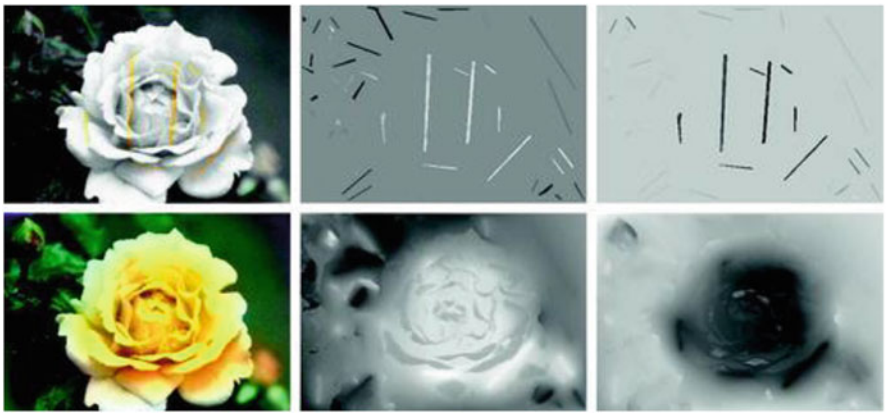


Fig. 5 Colorization experiment using the linear iteration of the neighborhood filter. *Top left*: input image with original luminance and initial data on the chromatic components. *Bottom right*: result image by applying the linear neighborhood scheme to the chromatic components using the initial chromatic data as boundary conditions. *Top middle and right*: initial data on the two chromatic components. *Bottom middle and bottom right*: final interpolated chromatic components

neighborhood filter process makes points tend to the local modes of the histogram but in a different framework: the first for images and the second for any dimensional clouds of points. This discontinuity or shock creation in regular zones of the image is not desirable for filtering or denoising applications. However, it can be used for image or video editing as proposed by Winnemoller et al. [73] in order to simplify video content and achieve a cartoon look.

Even if it may seem paradoxical, linear schemes have showed to be more useful than nonlinear ones for iterating the neighborhood filter; that is, the weight distribution for each pixel is computed once and is maintained during the whole iteration process. We will show in Sect. 4 that by computing the weights on an image and keeping them constant during the iteration process, a histogram concentration phenomenon makes the filter a powerful segmentation algorithm. The same iteration is useful to linearly diffuse or filter any initial data or seeds as proposed by Grady et al. [38] for medical image segmentation or [11] for colorization (see Fig. 5 for an example). The main hypothesis for this seed diffusion algorithm is that pixels having a similar gray level value should be related and are likely to belong to the same object. Thus, pixels of different sites are related as in a graph with a weight depending on the gray level distance. The iteration of the neighborhood filter on the graph is equivalent to the solution of the heat equation on the graph by taking the graph Laplacian. Eigenvalues and eigenvectors of such a graph Laplacian can be computed allowing the design of Wiener and thresholding filters on the graph (see [69] and [59, 60] for more details).

Both the neighborhood filter and the NL-means have been adapted and extended for other types of data and other image-processing tasks: for 3D data set points [17, 26, 35, 43, 79], and [42]; *demosaicking*, the operation which transforms the "R

or G or B" raw image in each camera into an "R and G and B" image [15, 51, 63]; *movie colorization*, [34] and [49]; *image inpainting* by proposing a nonlocal image inpainting variational framework with a unified treatment of geometry and texture [2] (see also [74]); *zooming* by a fractal-like technique where examples are taken from the image itself at different scales [29]; *movie flicker stabilization* [24], compensating spurious oscillations in the colors of successive frames; and *super-resolution*, an image zooming method by which several frames from a video, or several low-resolution photographs, can be fused into a larger image [62]. The main point of this super-resolution technique is that it gives up an explicit estimate of the motion, allowing actually for a multiple motion, since a block can look like several other blocks in the same frame. The very same observation is made in [30] for devising a super-resolution algorithm and in [22, 33].

2 Denoising

Analysis of Neighborhood Filter as a Denoising Algorithm

In this section, we will further investigate the neighborhood filter behavior as a denoising algorithm. We will consider the simplest neighborhood filter version which averages spatially close pixels with an intensity difference lower than a certain threshold h. By classical probability theory, the average of N random and i.i.d values has a variance N times smaller than the variance of the original values. However, this theoretical reduction is not observed when applying neighborhood filters.

In order to evaluate the noise reduction capability of the neighborhood filter, we apply it to a noise sample and evaluate the variance of the filtered sample. Let us suppose that we observe the realization of a white noise at a pixel $\mathbf{x}, n(\mathbf{x}) = a$. The nearby pixels with an intensity difference lower than h will be independent and identically distributed with the probability distribution function the restriction of the Gaussian to the interval $(a - h, a + h)$. If the research zone is large enough, then the average value will tend to the expectation of such a variable. Thus, the increase of the research zone and therefore of the number of pixels being averaged does not increase the noise reduction capability of the filter. Such a noise reduction factor is computed in the next result.

Theorem 1. *Assume that the $n(i)$ are i.i.d. with zero mean and variance σ^2. Then, the filtered noise by the neighborhood filter NF_h satisfies the following:*

- *The noise reduction depends only on the value of h,*

$$\mathrm{Var}\ \mathrm{NF}_h n(x) = f\left(\frac{h}{\sigma}\right)\sigma^2,$$

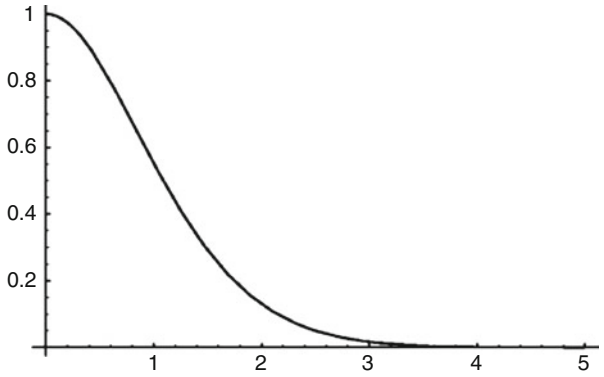


Fig. 6 Noise reduction function $f(x)$ given by Theorem 1

where

$$f(x) = \frac{1}{(2\pi)^{3/2}} \int_{\mathrm{R}} \frac{1}{\beta^2(a,x)} \left(e^{2xa} - 1\right)^2 e^{(a+x)^2} e^{\frac{-a^2}{2}} da$$

is a decreasing function with $f(0) = 1$ *and* $\lim_{x\to\infty} f(x) = 0$.

- *The values* $\mathrm{NF}_h n(\mathbf{x})$ *and* $\mathrm{NF}_h n(\mathbf{y})$ *are uncorrelated for* $\mathbf{x} \neq \mathbf{y}$.

The function $f(x)$ *is plotted in Fig. 6. The noise reduction increases as the ratio* h/σ *also does. We see that* $f(x)$ *is near zero for values of* x *over 2.5 or 3, that is, values of* h *over* 2.5σ *or* 3σ*, which justifies the values proposed in the original papers by Lee and Yaroslavsky. However, for a Gaussian variable, the probability of observing values at a distance of the average larger than 2.5 or 3 times the standard deviation is very small. Thus, by taking these values, we excessively increase the probability of mismatching pixels of different objects. Thus, close objects with an intensity contrast lower than* 3σ *will not be correctly denoised. This explains the decreasing performance of the neighborhood filter as the noise standard deviation increases.*

The previous theorem also tells us that the denoised noise values are still uncorrelated once the filter has been applied. This is easily justified since we showed that as the size ρ of the neighborhood increases, the filtered value tends to the expectation of the Gauss distribution restricted to the interval $(n(\mathbf{x}) - h, n(\mathbf{x}) + h)$. The filtered value is therefore a deterministic function of $n(\mathbf{x})$ and h. Independent random variables are mapped by a deterministic function on independent variables.

This property may seem anecdotic since noise is what we wish to get rid of. Now, it is impossible to totally remove noise. The question is how the remnants of noise look like. The transformation of a white noise into any correlated signal creates structure and artifacts. Only white noise is perceptually devoid of structure, as was pointed out by Attneave [3].

The only difference between the neighborhood filter and the bilateral or SUSAN filter is the way the spatial component is treated. While for the classical neighborhood all pixels within a certain distance are treated equally, for the bilateral filter, pixels closer to the reference pixel are more important. Even if this can seem a slight difference, this is crucial from a qualitative point of view, that is, the creation of artifacts.

It is easily shown that introducing the weighting function on the intensity difference instead of a non-weighted average does not modify the second property of Theorem 1, and the denoised noise values are still uncorrelated if ρ is large enough. However, the introduction of the spatial kernel by the bilateral or SUSAN filter affects this property. Indeed, the introduction of a spatial decay of the weights makes denoised values at close positions to be correlated.

There are two ways to show how denoising algorithms behave when they are applied to a noise sample. One of them is to find a mathematical proof that the pixels remain independent (or at least uncorrelated) and identically distributed random variables. The experimental device simply is to observe the effect of denoising on the simulated realization of a white noise. Figure 7 displays the filtered noises for the neighborhood filter, the bilateral filter, and other state-of-the-art denoising algorithms.

Neighborhood Filter Extension: The NL-Means Algorithm

Now in a very general sense inspired by the neighborhood filter, one can define as "neighborhood of a pixel **x**" any set of pixels **y** in the image such that a window around **y** looks like a window around **x**. All pixels in that neighborhood can be used for predicting the value at **x**, as was shown in [23,31] for texture synthesis and in [21, 80] for inpainting purposes. The fact that such a self-similarity exists is a regularity assumption, actually more general and more accurate than all regularity assumptions we consider when dealing with local smoothing filters, and it also generalizes a periodicity assumption of the image.

Let v be the noisy image observation defined on a bounded domain $\Omega \subset \mathbb{R}^2$, and let $\mathbf{x} \in \Omega$. The NL-means algorithm estimates the value of **x** as an average of the values of all the pixels whose Gaussian neighborhood looks like the neighborhood of **x**:

$$\mathrm{NL}(v)(x) = \frac{1}{C(x)} \int_{\Omega} e^{-\frac{\left(G_a * |v(x+.)-v(y+.)|^2\right)(0)}{h^2}} v(y)\, dy, \tag{3}$$

where G_a is a Gaussian kernel with standard deviation a, h acts as a filtering parameter, and $C(x) = \int_{\Omega} e^{-\frac{\left(G_a * |v(x+.)-v(z+.)|^2\right)(0)}{h^2}} dz$ is the normalizing factor. We recall that

$$\left(G_a * |v(x+.) - v(y+.)|^2\right)(0) = \int_{\mathrm{R}^2} G_a(t)\, |v(x+t) - v(y+t)|^2\, dt.$$

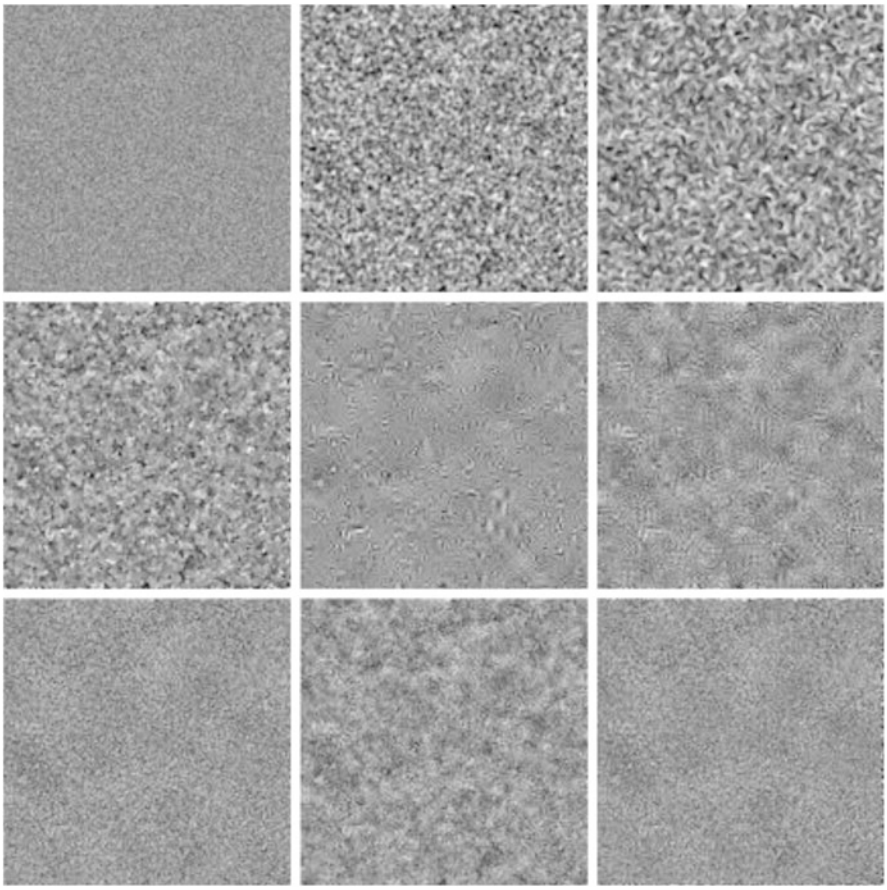


Fig. 7 The noise to noise criterion. From left to right and from top to bottom: original noise image of standard deviation 20, Gaussian convolution, anisotropic filtering, total variation, TIHWT, DCT empirical Wiener filter, neighborhood filter, bilateral filter, and the NL-means. Parameters have been fixed for each method so that the noise standard deviation is reduced by a factor 4. The filtered noise by the Gaussian filter and the total variation minimization are quite similar, even if the first one is totally blurred and the second one has created many high frequency details. The filtered noise by the hard wavelet thresholding looks like a constant image with superposed wavelets. The filtered noise by the neighborhood filter and the NL-means algorithm looks like a white noise. This is not the case for the bilateral filter, where low frequencies of noise are enhanced because of the spatial decay

We will see that the use of an entire window around the compared points makes this comparison more robust to noise. For the moment, we will compare the weighting distributions of both filters. Figure 8 illustrates how the NL-means algorithm chooses in each case a weight configuration adapted to the local geometry of the image. Then, the NL-means algorithm seems to provide a feasible and rational method to automatically take the best of all classical denoising algorithms, reducing for every possible geometric configuration the mismatched averaged pixels. It

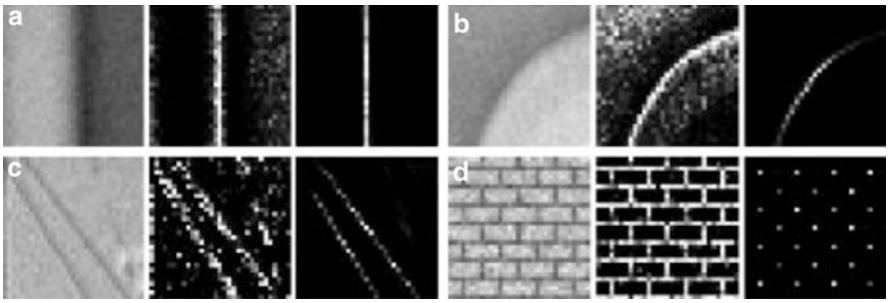


Fig. 8 Weight distribution of NL-means, the bilateral filter, and the anisotropic filter used to estimate the central pixel in four detail images. On the two right-hand-side images of each triplet, we display the weight distribution used to estimate the central pixel of the left image by the neighborhood and the NL-means algorithm. (**a**) In straight edges, the weights are distributed in the direction of the level line (as the mean curvature motion). (**b**) On curved edges, the weights favor pixels belonging to the same contour or level line, which is a strong improvement with respect to the mean curvature motion. In the cases of (**c**) and (**d**), the weights are distributed across the more similar configurations, even though they are far away from the observed pixel. This shows a behavior similar to a nonlocal neighborhood filter or to an ideal Wiener filter

preserves flat zones as the Gaussian convolution and straight edges as the anisotropic filtering while still restoring corners or curved edges and texture.

Due to the nature of the algorithm, one of the most favorable cases is the textural case. Texture images have a large redundancy. For each pixel, many similar samples can be found in the image with a very similar configuration, leading to a noise reduction and a preservation of the original image. In Fig. 9, one can see an example with a Brodatz texture. The Fourier transform of the noisy and restored images shows the ability of the algorithm to preserve the main features even in the case of high frequencies.

The NL-means seems to naturally extend the Gaussian, anisotropic, and neighborhood filtering. But it is not easily related to other state-of-the-art denoising methods as the total variation minimization [64], the wavelet thresholding [19, 27], or the local DCT empirical Wiener filters [76]. For this reason, we compare these methods visually in artificial denoising experiences (see [12] for a more comprehensive comparison).

Figure 10 illustrates the fact that a nonlocal algorithm is needed for the correct reconstruction of periodic images. Local smoothing filters and Wiener and thresholding methods are not able to reconstruct the wall pattern. Only NL-means and the global Fourier-Wiener filter reconstruct the original texture. The Fourier-Wiener filter is based on a global Fourier transform, which is able to capture the periodic structure of the image in a few coefficients. But this only is an ideal filter: the Fourier transform of the original image is being used. Figure 8d shows how NL-means chooses the correct weight configuration and explains the correct reconstruction of the wall pattern.

The NL-means algorithm is not only able to restore periodic or texture Images; natural images also have enough redundancy to be restored. For example, in a

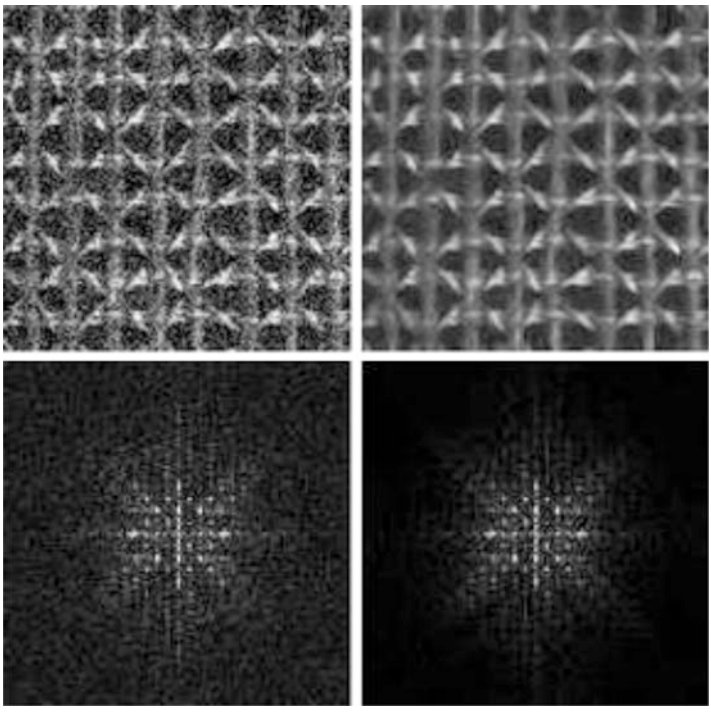


Fig. 9 NL-means denoising experiment with a Brodatz texture image. *Left*: noisy image with standard deviation 30. *Right*: NL-means restored image. The Fourier transforms of the noisy and restored images show how main features are preserved even at high frequencies

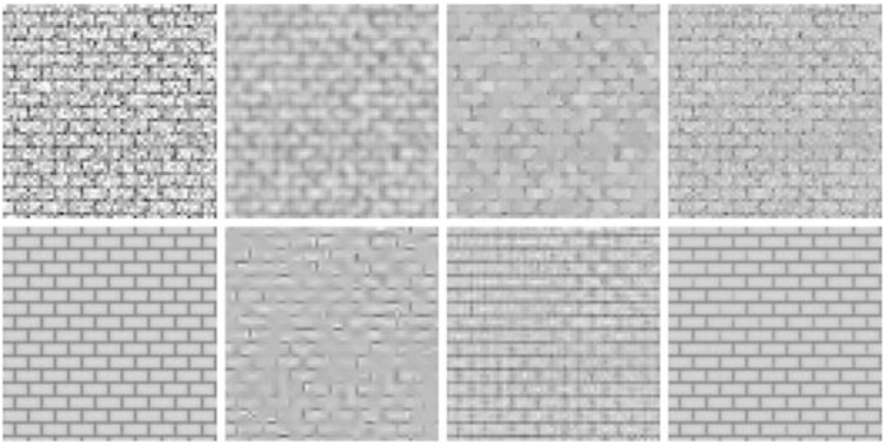


Fig. 10 Denoising experience on a periodic image. From left to right and from top to bottom: noisy image (standard deviation 35), Gauss filtering, total variation, neighborhood filter, Wiener filter (ideal filter), TIHWT (translation invariant hard thresholding), DCT empirical Wiener filtering, and NL-means

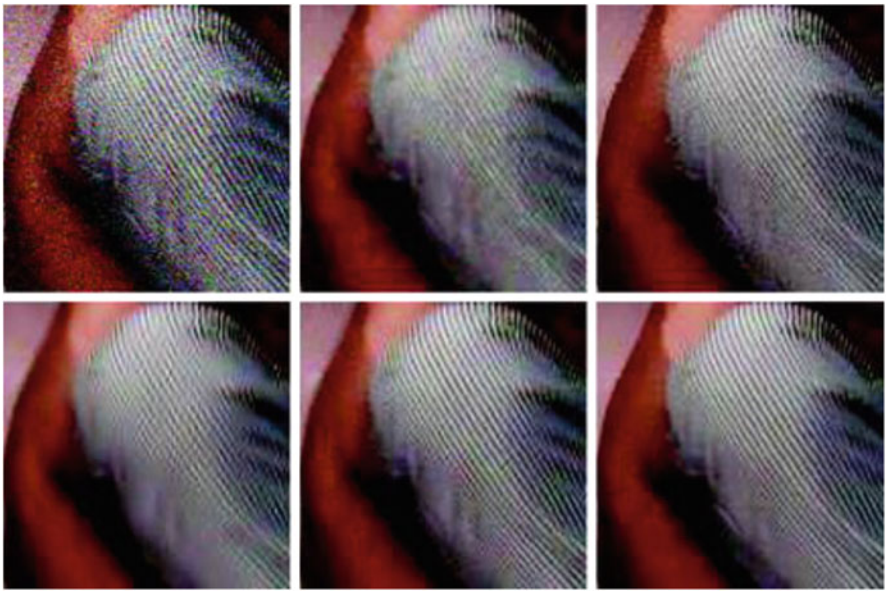


Fig. 11 Denoising experience on a natural image. From left to right and from top to bottom: noisy image (standard deviation 35), total variation, neighborhood filter, translation invariant hard thresholding (TIHWT), empirical Wiener, and NL-means

flat zone, one can find many pixels lying in the same region and with similar configurations. In a straight or curved edge, a complete line of pixels with a similar configuration is found. In addition, the redundancy of natural images allows us to find many similar configurations in faraway pixels.

Figure 11 shows that wavelet and DCT thresholding are well adapted to the recovery of oscillatory patterns. Although some artifacts are noticeable in both solutions, the stripes are well reconstructed. The DCT transform seems to be more adapted to this type of texture, and stripes are a little better reconstructed. For a much more detailed comparison between sliding window transform domain filtering methods and wavelet threshold methods, we refer the reader to [77]. NL-means also performs well on this type of texture, due to its high degree of redundancy.

The above description of movie denoising algorithms and its juxtaposition to the NL-means principle shows how the main problem, motion estimation, can be circumvented. In denoising, the more samples we have the happier we are. The *aperture problem* is just a name for the fact that there are many blocks in the next frame similar to a given one in the current frame. Thus, singling out one of them in the next frame to perform the motion compensation is an unnecessary and probably harmful step. A much simpler strategy that takes advantage of the aperture problem is to denoise a movie pixel by involving indiscriminately spatial and temporal similarities (see [14] for more details on this discussion). The algorithm favors pixels with a similar local configuration, as the similar configurations move, so do

the weights. Thus, the algorithm is able to follow the similar configurations when they move without any explicit motion computation (see Fig. 12).

Extension to Movies

Averaging filters are easily extended to the denoising of image sequences and video. The denoising algorithms involve indiscriminately pixels not belonging only to the same frame but also the previous and posterior ones.

In many cases, this straightforward extension cannot correctly deal with moving objects. For that reason, state-of-the-art movie filters are motion compensated (see [10] for a comprehensive review). The underlying idea is the existence of a "ground true" physical motion, which motion estimation algorithms should be able to estimate. Legitimate information should exist only along these physical trajectories. The *motion compensated filters* estimate explicitly the motion of the sequence by a motion estimation algorithm. The motion compensated movie yields a new stationary data on which an averaging filter can be applied. The motion compensated movie yields a new stationary data on which an averaging filter can be applied. The motion compensation neighborhood filter was proposed by Ozkan et al. [55]. We illustrate in Fig. 13 the improvement obtained with the proposed compensation.

One of the major difficulties in motion estimation is the ambiguity of trajectories, the so-called aperture problem. This problem is illustrated in Fig. 14. At most pixels, there are several options for the displacement vector. All of these options have a similar gray level value and a similar block around them. Now, motion estimators have to select one by some additional criterion.

3 Asymptotic

PDE Models and Local Smoothing Filters

According to Shannon's theory, a signal can be correctly represented by a discrete set of values, the "samples," only if it has been previously smoothed. Let us start with u_0 the physical image, a real function defined on a bounded domain $\Omega \subset \mathbb{R}^2$. Then a blur optical kernel k is applied, i.e., u_0 is convolved with k to obtain an observable signal $k^* u_0$. Gabor remarked in 1960 that the difference between the original and the blurred images is roughly proportional to its Laplacian, $\Delta u = u_{xx} + u_{yy}$. In order to formalize this remark, we have to notice that k is spatially concentrated and that we may introduce a scale parameter for k, namely, $k_h(x) = h^{-1}k\left(h^{-\frac{1}{2}}x\right)$. If, for instance, u is C^2 and bounded and if k is a radial function in the Schwartz class, then

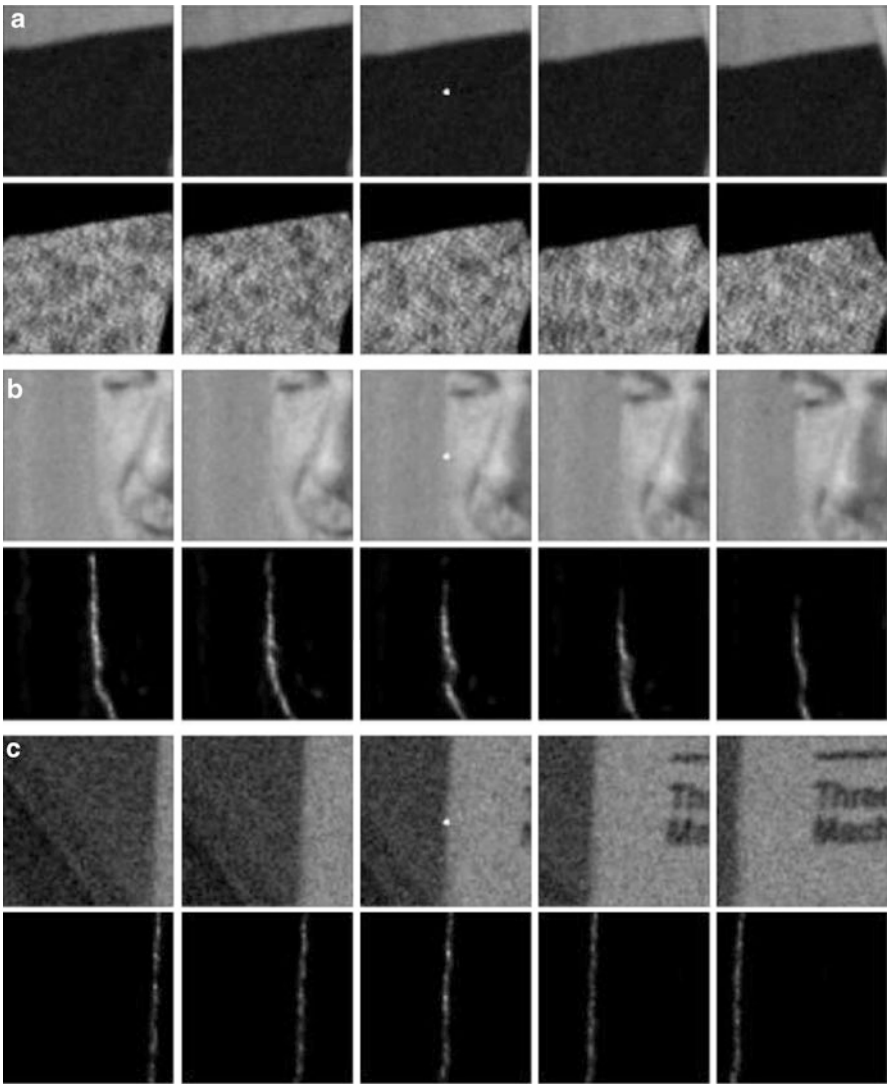


Fig. 12 Weight distribution of NL-means applied to a movie. In (**a**), (**b**), and (**c**), the first row shows a five frames image sequence. In the second row, the weight distribution used to estimate the central pixel (in *white*) of the middle frame is shown. The weights are equally distributed over the successive frames, including the current one. They actually involve all the candidates for the motion estimation instead of picking just one per frame. The aperture problem can be taken advantage of for a better denoising performance by involving more pixels in the average Fig. 7 displays the application of the denoising methods to a white noise. We display the filtered noise

Fig. 13 Comparison of static filters, motion compensated filters, and NL-means applied to an image sequence. *Top*: three frames of the sequence are displayed. *Middle* and *left to right*: neighborhood filter, motion compensated neighborhood filter, and the NL-means. (AWA). *Bottom*: the noise removed by each method (difference between the noisy and filtered frame). Motion compensation improves the static algorithms by better preserving the details and creating less blur. We can read the titles of the books in the noise removed by AWA. Therefore, that much information has been removed from the original. Finally, the NL-means algorithm (*bottom row*) has almost no noticeable structure in its removed noise. As a consequence, the filtered sequence has kept more details and is less blurred

$$\frac{u_{0*}k_h(x) - u_0(x)}{h} \to c\Delta u_0(x).$$

Hence, when h gets smaller, the blur process looks more and more like the heat equation

$$u_t = c\Delta u, \;\; u(0) - u_0.$$

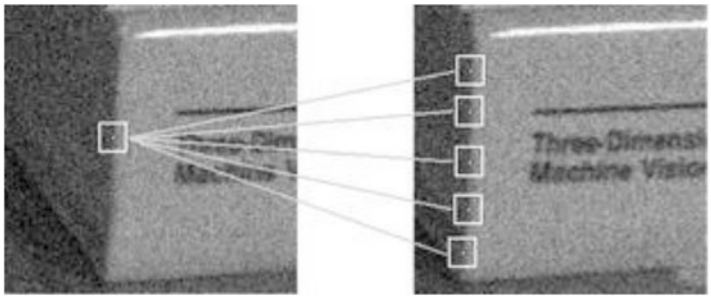


Fig. 14 Aperture problem and the ambiguity of trajectories are the most difficult problems in motion estimation: There can be many good matches. The motion estimation algorithms must pick one

Thus, Gabor established a first relationship between local smoothing operators and PDEs. The classical choice for k is the Gaussian kernel.

Remarking that the optical blur is equivalent to one step of the heat equation, Gabor deduced that we can, to some extent, deblur an image by reversing the time in the heat equation, $u_t = -\Delta u$. Numerically, this amounts to subtracting the filtered version from the original image:

$$u - G_{h*}u = -h^2 \Delta u + o(h^2).$$

This leads to considering the reverse heat equation as an image restoration, ill-posed though it is. The time-reversed heat equation was stabilized in the Osher-Rudin shock filter [54] who proposed

$$u_t = -\text{sign}(\mathcal{L}(u)) \, |Du| \, ,$$

where the propagation term $|Du|$ is tuned by the sign of an edge detector $\Lambda(u)$. The function $\Lambda(u)$ changes sign across the edges where the sharpening effect therefore occurs. In practice, $\Lambda(u) = \Delta u$ and the equation is related to a reverse heat equation.

The early Perona-Malik "anisotropic diffusion" [57] is directly inspired from the Gabor remark. It reads

$$u_t = \text{div}\left(g\left(|Du|^2\right) Du\right),$$

where $g : [0, +\infty) \to [0, +\infty)$ is a smooth decreasing function satisfying $g(0) = 1, \lim_{s \to +\infty} g(s) = 0$. This model is actually related to the preceding ones. Let us consider the second derivatives of u in the directions of Du and $Du^{\perp}$:

$$u_{\eta\eta} = D^2 u \left(\frac{Du}{|Du|}, \frac{Du}{|Du|} \right), \quad u_{\xi\xi} = D^2 u \left(\frac{Du^{\perp}}{|Du|}, \frac{Du^{\perp}}{|Du|} \right).$$

Then, Eq. (5) can be rewritten as

$$u_t = g\left(|Du|^2\right) u_{\xi\xi} + h\left(|Du|^2\right) u_{\eta\eta},$$

where $h(s) = g(s) + 2sg'(s)$. Perona and Malik proposed the function $g(s) = \frac{1}{1+s/k}$. In this case, the coefficient of the first term is always positive and this term therefore appears as a one-dimensional diffusion term in the orthogonal direction to the gradient. The sign of the second coefficient, however, depends on the value of the gradient. When $|Du|^2 < k$, this second term appears as a one-dimensional diffusion in the gradient direction. It leads to a reverse heat equation term when $|Du|^2 > k$.

The Perona-Malik model has got many variants and extensions. Tannenbaum and Zucker [45] proposed, endowed in a more general shape analysis framework, the simplest equation of the list:

$$u_t = |Du| \operatorname{div}\left(\frac{Du}{|Du|}\right) = u_{\xi\xi}.$$

This equation had been proposed some time before in another context by Sethian [66] as a tool for front propagation algorithms. This equation is a "pure" diffusion in the direction orthogonal to the gradient and is equivalent to the anisotropic filter AF [40]:

$$\mathrm{AF}_h u(x) = \int G_h(t) u(x + t\xi) dt,$$

where $\xi = Du(\mathbf{x})^{\perp}/|Du(\mathbf{x})|$ and $G_h(t)$ denotes the one-dimensional Gauss function with variance h^2.

This diffusion is also related to two models proposed in image restoration. The Rudin-Osher-Fatemi [64] total variation model leads to the minimization of the total variation of the image $TV(u) = \int |Du|$, subject to some constraints. The steepest descent of this energy reads, at least formally,

$$\frac{\partial u}{\partial t} = \operatorname{div}\left(\frac{Du}{|Du|}\right) \tag{4}$$

which is related to the mean curvature motion and to the Perona-Malik equation when $g\left(|Du|^2\right) = \frac{1}{|Du|}$. This particular case, which is not considered in [57], yields again (4). An existence and uniqueness theory is available for this equation [1].

Asymptotic Behavior of Neighborhood Filters (Dimension 1)

Let u denote a one-dimensional signal defined on an interval $I \subset \mathbb{R}$ and consider the neighborhood filter

$$\mathrm{NF}_{h,\rho}u(x) = \frac{1}{C(x)} \int_{x-\rho}^{x+\rho} u(y) e^{-\frac{|u(y)-u(x)|^2}{h^2}} dy,$$

where $C(x) = \int_{x-\rho}^{x+\rho} e^{-\frac{|u(y)-u(x)|^2}{h^2}} dy$.

The following theorem describes the asymptotical behavior of the neighborhood filter in 1D. The proof of this theorem and next ones in this section can be found in [13].

Theorem 2. *Suppose $u \in C^2(I)$, and let $\rho, h, \alpha > 0$ such that $\rho, h \to 0$ and $h = 0(\rho^\alpha)$. Consider the continuous function $g(t) = \frac{te^{-t^2}}{E(t)}$, for $t \neq 0, g(0) = \frac{1}{2}$, where $E(t) = 2\int_0^t e^{-s^2} ds$. Let f be the continuous function*

$$f(t) = \frac{g(t)}{t^2} + g(t) - \frac{1}{2t^2}, \quad f(0) = \frac{1}{6}.$$

Then, for $x \in \mathbb{R}$,

1. *If $\alpha < 1, \mathrm{NF}_{h,\rho}u(x) - u(x) \simeq \frac{u''(x)}{6}\rho^2$.*
2. *If $\alpha = 1, \mathrm{NF}_{h,\rho}u(x) - u(x) \simeq f\left(\frac{\rho}{h}|u'(x)|\right) u''(x)\rho^2$.*
3. *If $1 < \alpha < \frac{3}{2}, \mathrm{NF}_{h,\rho}u(x) - u(x) \simeq g\left(\rho^{1-\alpha}|u'(x)|\right) u''(x)\rho^2$.*

According to Theorem 2, the neighborhood filter makes the signal evolve proportionally to its second derivative. The equation $u_t = cu'$ acts as a smoothing or enhancing model depending on the sign of c. Following the previous theorem, we can distinguish three cases depending on the values of h and ρ. First, if h is much larger than ρ, the second derivative is weighted by a positive constant and the signal is therefore filtered by a heat equation. Second, if h and ρ have the same order, the sign and magnitude of the weight is given by $f\left(\frac{\rho}{h}|u'(x)|\right)$. As the function f takes positive and negative values (see Fig. 15), the filter behaves as a filtering/enhancing algorithm depending on the magnitude of $|u'(x)|$. If B denotes the zero of f, then a filtering model is applied wherever $|u'| < B\frac{h}{\rho}$ and an enhancing model wherever $|u'| > B\frac{h}{\rho}$. The intensity of the enhancement tends to zero when the derivative tends to infinity. Thus, points x where $|u'(x)|$ is large are not altered. The transition of the filtering to the enhancement model creates a singularity in the filtered signal. In the last case, ρ is much larger than h and the sign and magnitude of the weight is given by $g\left(\frac{\rho}{h}|u'(x)|\right)$. Function g is positive and decreases to zero. If the derivative of u is bounded, then $\frac{\rho}{h}|u'(x)|$ tends to infinity and the intensity of the filtering to zero. In this case, the signal is hardly modified.

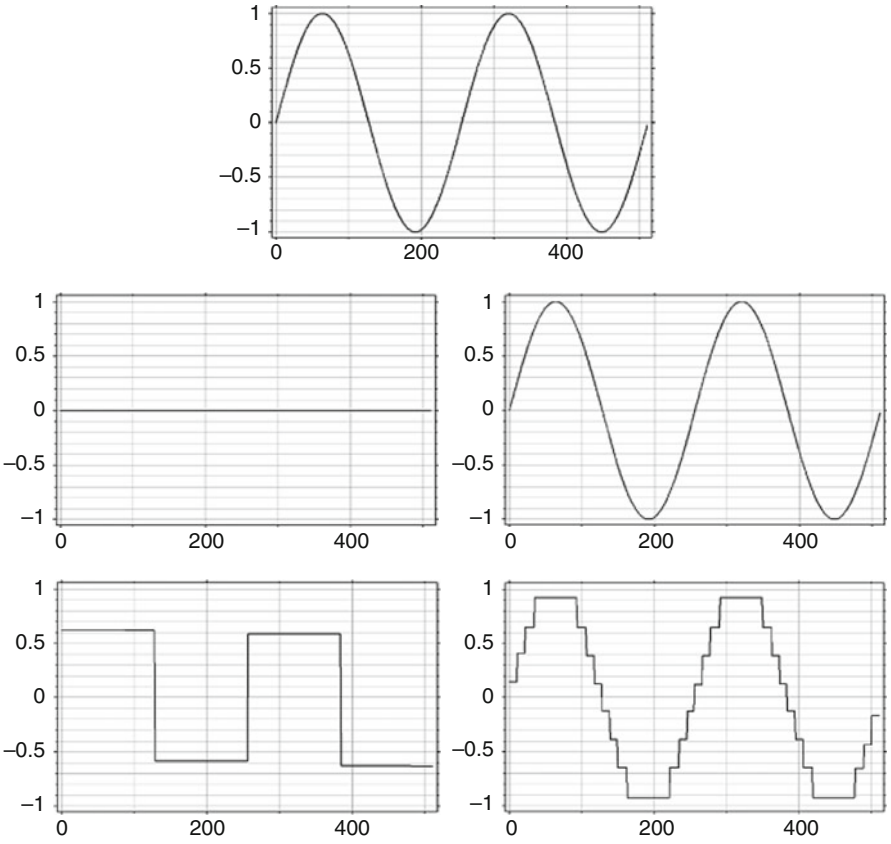


Fig. 15 One dimensional neighborhood filter experiment. The neighborhood filter is iterated until the steady state is attained for different values of the ratio ρ/h. *Top*: Original sine signal. *Middle left*: filtered signal with $\rho/h = 10^{-8}$. *Middle right*: filtered signal with $\rho/h = 10^{8}$. *Bottom left*: filtered signal with $\rho/h = 2$. *Bottom right*: filtered signal with $\rho/h = 5$. The examples corroborate the results of Theorem 2. If ρ/h tends to zero the algorithm behaves like a heat equation and the filtered signal tends to a constant. If, instead, ρ/h tends to infinity the signal is hardly modified. If ρ and h have the same order, the algorithm presents a filtering/enhancing dynamic. Singularities are created due to the transition of smoothing to enhancement. The number of enhanced regions strongly depends upon the ratio $\frac{\rho}{h}$ as illustrated in the *bottom figures*

In summary, a neighborhood filter in dimension 1 shows interesting behavior only if ρ and h have the same order of magnitude, in which case the neighborhood filter behaves like a Perona-Malik equation. It enhances edges with a gradient above a certain threshold and smoothes the rest.

Figure 16 illustrates the behavior of the one-dimensional neighborhood filter. The algorithm is iterated until the steady state is attained on a sine signal for different values of the ratio ρ/h. The results of the experiment corroborate the asymptotical expansion of Theorem 2. In the first experiment, $\rho/h = 10^{-8}$ and the neighborhood

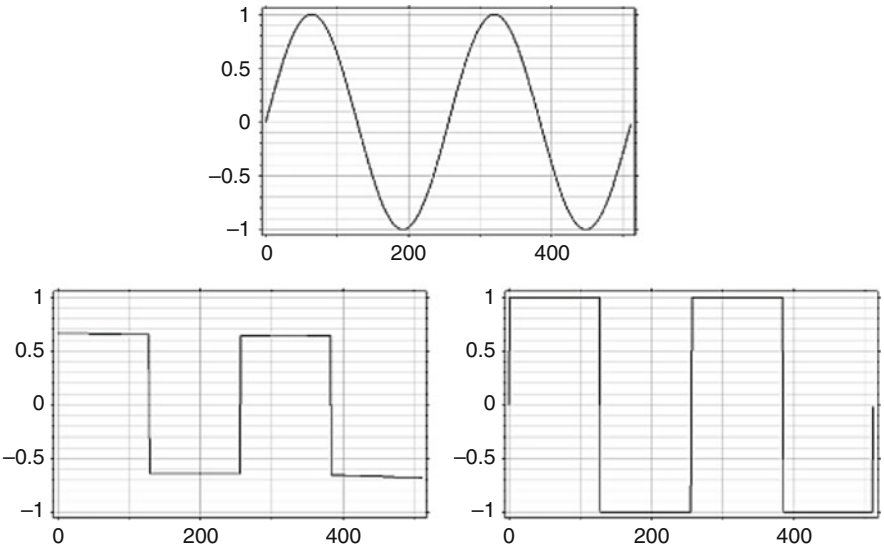


Fig. 16 One-dimensional neighborhood filter experiment. The neighborhood filter is iterated until the steady state is attained for different values of the ratio ρ/h. *Top*: original sine signal. *Middle left*: filtered signal with $\rho/h = 10^{-8}$. *Middle right*: filtered signal with $\rho/h = 10^{8}$. *Bottom left*: filtered signal with $\rho/h = 2$. *Bottom right*: filtered signal with $\rho/h = 5$. The examples corroborate the results of Theorem 2. If ρ/h tends to zero, the algorithm behaves like a heat equation and the filtered signal tends to a constant. If, instead, ρ/h tends to infinity, the signal is hardly modified. If ρ and h have the same order, the algorithm presents a filtering/enhancing dynamic. Singularities are created due to the transition of smoothing to enhancement. The number of enhanced regions strongly depends upon the ratio $\frac{\rho}{h}$, as illustrated in the bottom figures

filter is equivalent to a heat equation. The filtered signal tends to a constant. In the second experiment, $\rho/h = 10^8$ and the value $g\left(\frac{\rho}{h}|u'|\right)$ is nearly zero. As predicted by the theorem, the filtered signal is nearly identical to the original one. The last two experiments illustrate the filtering/enhancing behavior of the algorithm when h and ρ have similar values. As predicted, an enhancing model is applied where the derivative is large. Many singularities are being created because of the transition of the filtering to the enhancing model. Unfortunately, the number of singularities and their position depend upon the value of ρ/h. This behavior is explained by Theorem 2(2). Figure 22 illustrates the same effect in the 2D case.

The filtering/enhancing character of the neighborhood filter is very different from a pure enhancing algorithm like the Osher-Rudin shock filter. Figures 17 and 18 illustrate these differences. In Fig. 17, the minimum and the maximum of the signal have been preserved by the shock filter, while these two values have been significantly reduced by the neighborhood filter. This filtering/enhancing effect is optimal when the signal is noisy. Figure 18 shows how the shock filter creates artificial steps due to the fluctuations of noise, while the neighborhood filter reduces the noise avoiding any spurious shock. Parameter h has been chosen larger than

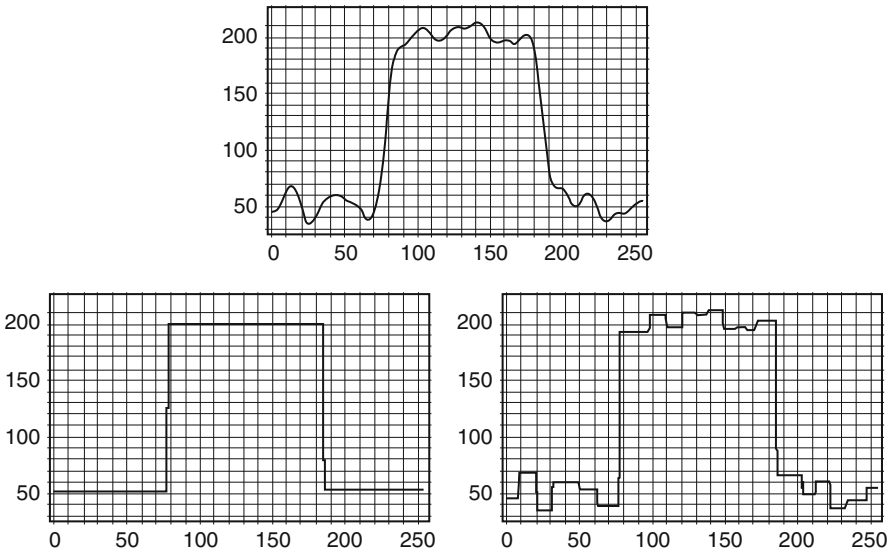


Fig. 17 Comparison between the neighborhood filter and the shock filter. *Top*: original signal. *Bottom left*: application of the neighborhood filter. *Bottom right*: application of the shock filter. The minimum and the maximum of the signal have been preserved by the shock filter and reduced by the neighborhood filter. This fact illustrates the filtering/enhancing character of the neighborhood filter compared with a pure enhancing filter

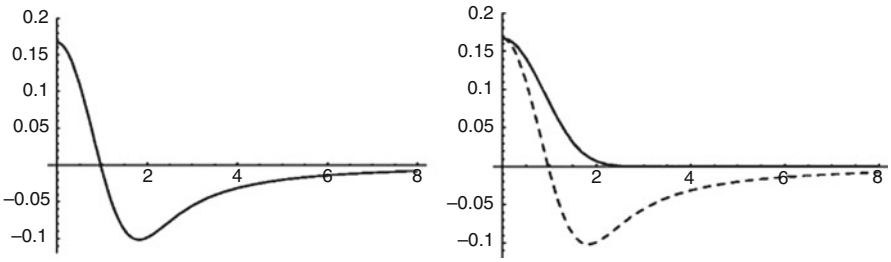


Fig. 18 Comparison between the neighborhood filter and the shock filter. *Top*: original signal. *Bottom left*: application of the neighborhood filter. *Bottom right*: application of the shock filter. The shock filter is sensitive to noise and creates spurious steps. The filtering/enhancing character of the neighborhood filter avoids this effect

the amplitude of noise in order to remove it. Choosing an intermediate value of h, artificial steps could also be generated on points where the noise amplitude is above this parameter value.

The Two-Dimensional Case

The following theorem extends the previous result to the two-dimensional case.

Theorem 3. *Suppose $u \in C^2(\Omega)$, and let $\rho, h, \alpha > 0$ such that $\rho, h \to 0$ and $h = O(\rho^\alpha)$. Let us consider the continuous function $\tilde{g}$ defined by $\tilde{g}(t) = \frac{1}{3}\frac{te^{-t^2}}{\frac{3}{4}E(t)}$, for $t \neq 0, \tilde{g}(0) = \frac{1}{6}$, where $E(t) = 2\int_0^t e^{-s^2} ds$. Let $\tilde{f}$ be the continuous function defined by*

$$\tilde{f}(t) = 3\tilde{g}(t) + \frac{3\tilde{g}(t)}{t^2} - \frac{1}{2t^2}, \quad \tilde{f}(0) = \frac{1}{6}.$$

Then, for $\mathbf{x} \in \Omega$,

1. *If $\alpha < 1$,*

$$\mathrm{NF}_{h,\rho}u(x) - u(x) \simeq \frac{\Delta u(x)}{6}\rho^2.$$

2. *If $\alpha = 1$,*

$$\mathrm{NF}_{h,\rho}u(x) - u(x) \simeq \left[\tilde{g}\left(\frac{\rho}{h}\left|Du(x)\right|\right)u_{\xi\xi}(x) + \tilde{f}\left(\frac{\rho}{h}\left|Du(x)\right|\right)u_{\eta\eta}(x)\right]\rho^2.$$

3. *If $1 < \alpha < \frac{3}{2}$,*

$$\mathrm{NF}_{h,\rho}u(x) - u(x) \simeq \tilde{g}\left(\rho^{1-\alpha}\left|Du(x)\right|\right)\left[u_{\xi\xi}(x) + 3u_{\eta\eta}(x)\right]\rho^2.$$

where $\xi = Du(\mathbf{x})^\perp/|Du(\mathbf{x})|$ and $\eta = Du(\mathbf{x})/|Du(\mathbf{x})|$.

According to Theorem 3, the two-dimensional neighborhood filter acts as an evolution PDE with two terms. The first term is proportional to the second derivative of u in the direction $\xi = Du(\mathbf{x})^\perp/|Du(\mathbf{x})|$, which is tangent to the level line passing through $\mathbf{x}$. The second term is proportional to the second derivative of u in the direction $\eta = Du(\mathbf{x})/|Du(\mathbf{x})|$, which is orthogonal to the level line passing through $\mathbf{x}$. Like in the one-dimensional case, the evolution equations $u_t = c_1 u_{\xi\xi}$ and $u_t = c_2 u_{\eta\eta}$ act as filtering or enhancing models depending on the signs of c_1 and c_2. Following the previous theorem, we can distinguish three cases, depending on the values of h and ρ.

First, if h is much larger than ρ, both second derivatives are weighted by the same positive constant. Thus, the sum of both terms is equivalent to the Laplacian of u, Δu, and we get back to Gaussian filtering.

Second, if h and ρ have the same order of magnitude, the neighborhood filter behaves as a filtering/enhancing algorithm. The coefficient of the diffusion in the tangential direction, $u_{\xi\xi}$, is given by $\tilde{g}\left(\frac{\rho}{h}|Du|\right)$. The function $\tilde{g}$ is positive and

decreasing. Thus, there is always diffusion in that direction. The weight of the normal diffusion, $u_{\eta\eta}$, is given by $\tilde{f}\left(\frac{\rho}{h}|Du|\right)$. As the function $\tilde{f}$ takes positive and negative values (see Fig. 15), the filter behaves as a filtering/enhancing algorithm in the normal direction and depending on $|Du|$. If $\tilde{B}$ denotes the zero of $\tilde{f}$, then a filtering model is applied wherever $|Du| < \tilde{B}\frac{h}{\rho}$ and an enhancing strategy wherever $|Du| > \tilde{B}\frac{h}{\rho}$. The intensity of the filtering in the tangent diffusion and the enhancing in the normal diffusion tend to zero when the gradient tends to infinity. Thus, points with a very large gradient are not altered.

Finally, if ρ is much larger than h, the value $\frac{\rho}{h}$ tends to infinity and then the filtering magnitude $\tilde{g}\left(\frac{\rho}{h}|Du|\right)$ tends to zero. Thus, the original image is hardly altered. Let us mention that similar calculations were performed in a particular case for the neighborhood median filter by Masnou [52].

We observe that when ρ and h have the same order, the neighborhood filter asymptotically behaves like a Perona-Malik model. Let us be more specific about this comparison. Taking $g(s) = \tilde{g}\left(s^{\frac{1}{2}}\right)$ in the Perona-Malik Eq. (6), we obtain

$$u_t = \tilde{g}\left(|Du|\right) u_{\xi\xi} + \tilde{h}\left(|Du|\right) u_{\eta\eta}, \tag{5}$$

where $\tilde{h}(s) = \tilde{g}(s) + s\tilde{g}'(s)$. Thus, the Perona-Malik model and the neighborhood filter can be decomposed in the same way and with exactly the same weight in the tangent direction. Then the function $\tilde{h}$ has the same behavior as $\tilde{f}$ (Theorem 3), as can be observed in Fig. 19. Thus, in this case, a neighborhood filter has the same qualitative behavior as a Perona-Malik model, even if we cannot rewrite it exactly as such.

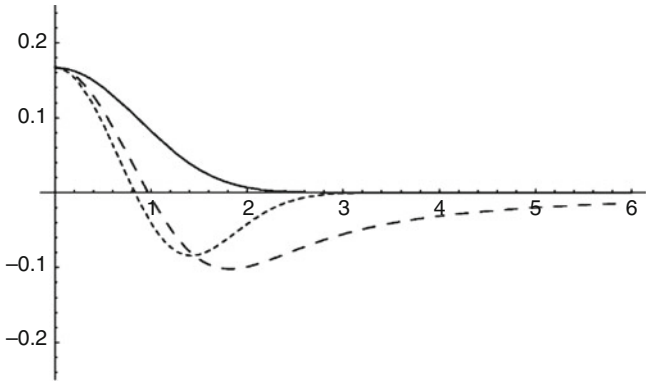


Fig. 19 Weight comparison of the neighborhood filter and the Perona-Malik equation. Magnitude of the tangent diffusion (*continuous line*, identical for both models) and normal diffusion (*dashed line- -*) of Theorem 3. Magnitude of the tangent diffusion (*continuous line*) and normal diffusion (*dashed line- - -*) of the Perona-Malik model (5). Both models show nearly the same behavior

Figure 22 displays a comparison of the neighborhood filter and the Perona-Malik model. We display a natural image and the filtered images by both models. These solutions have a similar visual quality and tend to display flat zones and artificial contours inside the smooth regions. Figure 23 corroborates this visual impression. We display the level lines of both filtered solutions. As expected from the above consistency theorems, for both models the level lines of the original image tend to concentrate, thus creating large flat zones separated by edges. The solutions are very close, up to the obvious very different implementations. The neighborhood filter is implemented exactly as in its definition and the Perona-Malik model by the explicit difference scheme proposed in the original paper.

A Regression Correction of the Neighborhood Filter

In the previous sections, we have shown the enhancing character of the neighborhood filter. We have seen that the neighborhood filter, like the Perona-Malik model, can create large flat zones and spurious contours inside smooth regions. This effect depends upon a gradient threshold which is hard to fix in such a way as to always separate the visually smooth regions from edge regions. In order to avoid this undesirable effect, let us analyze in more detail what happens with the neighborhood filter in the one-dimensional case.

Figure 20 shows a simple illustration of this effect. For each x in the convex part of the signal, the filtered value is the average of the points y such that $u(x) - h < u(y) < u(x) + h$ for a certain threshold h. As it is illustrated in the figure, the number of points satisfying $u(x) - h < u(y) \leq u(x)$ is larger than the number of points satisfying $u(x) \leq u(y) < u(x) + h$. Thus, the average value $\mathrm{YNF}(x)$ is smaller than $u(x)$, enhancing this part of the signal. A similar argument can be applied in the concave parts of the signal, dealing with the same enhancing effect. Therefore, shocks will be created inside smooth zones where concave and convex parts meet. Figure 20 also shows how the mean is not a good estimate of $u(x)$ in this case. In the same figure, we display the regression line approximating u

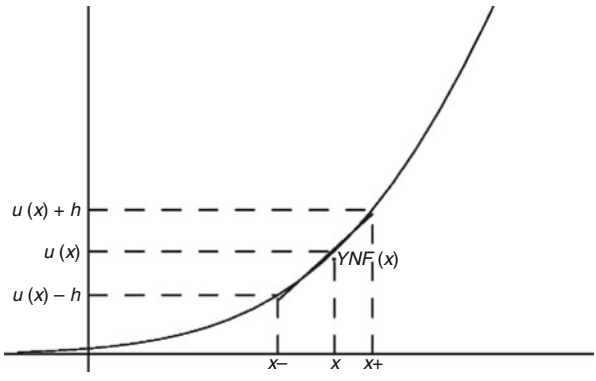


Fig. 20 Illustration of the shock effect of the YNF on the convex of a signal. The number of points y satisfying $u(x) - h < u(y) \leq u(x)$ is larger than the number satisfying $u(x) \leq u(y) < u(x) + h$. Thus, the average value $\mathrm{YNF}(x)$ is smaller than $u(x)$, enhancing that part of the signal. The regression line of u inside (x_-, x_+) better approximates the signal at x

inside $(u^{-1}(u(x)-h), u^{-1}(u(x)+h))$. We see how the value of the regression line at x better approximates the signal. In the sequel, we propose to correct the neighborhood filter with this better estimate.

In the general case, this linear regression strategy amounts to finding for every point $\mathbf{x}$ the plane locally approximating u in the following sense:

$$\min_{a_0,a_1} \int_{B_\rho(x)} w(x,y)(u(y)-a_1y_1-a_0)^2 dy, \qquad w(x,y) = e^{-\frac{|u(y)-|}{h^2}}$$

and then replacing $u(\mathbf{x})$ by the filtered value $a_1x_1 + a_0$. The weights used to define the minimization problem are the same as the ones used by the neighborhood filter. Thus, the points with a gray level value close to $u(x)$ will have a larger influence in the minimization process than those with a further gray level value. We denote the above linear regression correction by $\mathrm{LNF}_{h,\rho}$. Taking $a_1 = 0$ and then approximating u by a constant function, the minimization (5) goes back to the neighborhood filter.

This minimization was originally proposed by Cleveland [18] with a weight family not depending on the function u but only on the spatial distance of $\mathbf{x}$ and $\mathbf{y}$. A similar scheme incorporating u in the weight computation has been statistically studied in [61]. The authors propose an iterative procedure that describes for every point the largest possible neighborhood in which the initial data can be well approximated by a parametric function.

Another similar strategy is the interpolation by ENO schemes [41]. The goal of ENO interpolation is to obtain a better adapted prediction near the singularities of the data. For each point it selects different stencils of fixed size M and for each stencil reconstructs the associated interpolation polynomial of degree M. Then the *least oscillatory* polynomial is selected by some prescribed numerical criterion. The selected stencils tend to escape from large gradients and discontinuities.

The regression strategy also tends to select the right points in order to approximate the function. Instead of choosing a certain interval, all the points are used in the polynomial reconstruction, but weighted by the gray level differences.

As in the previous sections, let us analyze the asymptotic behavior of the linear regression correction. We compute the asymptotic expansion of the filter when $0 < \alpha \leq 1$. We showed that when $\alpha > 1$, the signal is hardly modified.

For the sake of completeness, we first compute the asymptotic expansion in the one-dimensional case.

Theorem 4. *Suppose $u \in C^2(I)$, and let $\rho, h, \alpha > 0$ such that $\rho, h \to 0$ and $h = O(\rho^\alpha)$. Let $\tilde{f}$ be the continuous function defined as $\tilde{f}(0) = \frac{1}{6}$,*

$$\tilde{f}(t) = \frac{1}{4t^2}\left(1 - \frac{2te^{-t^2}}{E(t)}\right),$$

for $t \neq 0$, where $E(t) = 2\int_0^t e^{-s^2} ds$. Then, for $x \in \mathbb{R}$,

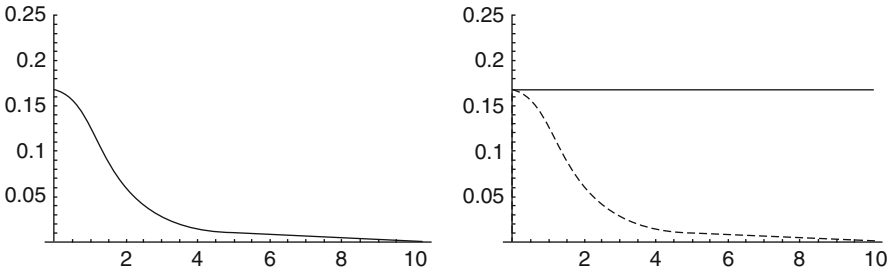


Fig. 21 Weighting functions of Theorems 4 and 5. *Left*: function $\tilde{f}$ of Theorem 4. *Right*: constant function 1/6 (*continuous line*) and function $\tilde{f}$ (*dashed line*) of Theorem 5

1. *If* $\alpha < 1, \mathrm{LNF}_{h,\rho}u(x) - u(x) \simeq \frac{u''(x)}{6}\rho^2$.
2. *If* $\alpha = 1, \mathrm{NF}_{h,\rho}u(x) - u(x) \simeq \tilde{f}\left(\frac{\rho}{h}|u'(x)|\right)u''(x)\rho^2$.

Theorem 4 shows that the $\mathrm{LNF}_{h,\rho}$ filter lets the signal evolve proportionally to its second derivative, as the neighborhood filter does. When h is larger than ρ, the filter is equivalent to the original neighborhood filter and the signal is filtered by a heat equation. When ρ and h have the same order, the sign and magnitude of the filtering process is given by $\tilde{f}\left(\frac{\rho}{h}|u'(x)|\right)$ (see Fig. 21). This function is positive and quickly decreases to zero. Thus, the signal is filtered by a heat equation of decreasing magnitude and is not altered wherever the derivative is very large.

The same asymptotic expansion can be computed in the two-dimensional case.

Theorem 5. *Suppose* $u \in C^2(\Omega)$, *and let* $\rho, h, \alpha > 0$ *such that* $\rho, h \to 0$ *and* $h = O(\rho^\alpha)$. *Let* $\tilde{f}$ *be the continuous function defined as* $\tilde{f}(0) = \frac{1}{6}$,

$$\tilde{f}(t) = \frac{1}{4t^2}\left(1 - \frac{2te^{-t^2}}{E(t)}\right),$$

for $t \neq 0$, *where* $E(t) = 2\int_0^t e^{-s^2}ds$. *Then, for* $\mathbf{x} \in \Omega$,

1. *If* $\alpha < 1$,

$$\mathrm{LNF}_{h,\rho}u(x) - u(x) \simeq \frac{\Delta u(x)}{6}\rho^2.$$

2. *If* $\alpha = 1$,

$$\mathrm{LNF}_{h,\rho}u(x) - u(x) \simeq \left[\tilde{f}\left(\frac{\rho}{h}|Du(x)|\right)u_{\eta\eta}(x)(x) + \frac{1}{6}u_{\xi\xi}(x)\right]\rho^2.$$

According to the previous theorem, the filter can be written as the sum of two diffusion terms in the direction of ξ and η. When h is much larger than ρ, the linear regression correction is equivalent to the heat equation like the original neighborhood filter. When ρ and h have the same order, the behavior of the linear regression algorithm is very different from the original neighborhood filter. The function weighting the tangent diffusion is a positive constant. The function weighting the normal diffusion is positive and decreasing (see Fig. 21), and therefore there is no enhancing effect. The algorithm combines the tangent and normal diffusion wherever the gradient is small. Wherever the gradient is larger, the normal diffusion is canceled and the image is filtered only in its tangent direction. This subjacent PDE was already proposed as a diffusion equation in [4]. This diffusion makes the level lines evolve proportionally to their curvature. In the Perona-Malik model, the diffusion is stopped near the edges. In this case, the edges are filtered by a mean curvature motion.

It may be asked whether the modified neighborhood filter still preserves signal discontinuities. The answer is yes. It is easily checked that for small enough h, all piecewise affine functions with smooth jump curves are steady. Thus, the behavior is the same as for the classical neighborhood filter. Our asymptotic analysis is of course not valid for such functions, but only for smooth functions.

As a numerical scheme, the linear regression neighborhood filter allows the implementation of a mean curvature motion without the computation of gradients and orientations. When the gradient is small, the linear regression filter naturally behaves like the heat equation. This effect is introduced on typical schemes implementing the mean curvature motion. In flat zones, the gradient is not well defined and some kind of isotropic diffusion must be applied. Therefore, the linear regression neighborhood filter naturally extends the mean curvature motion and yields a stable numerical scheme for its computation, independent of gradient orientations.

Figure 22 displays an experiment comparing the $\mathrm{LNF}_{h,\rho}$ with the neighborhood filter and the Perona-Malik equation. The linear correction does not create any contour or flat zone inside the smooth regions. Figure 23 displays the level lines of the previous experiment. The level lines of the $\mathrm{LNF}_{h,\rho}$ are filtered by a mean curvature motion, and they do not get grouped creating flat zones. The same effect is illustrated in Fig. 24.

The Vector-Valued Case

Let u be a vector-valued function defined on a bounded domain $\Omega \subset \mathbb{R}^2$, $u : \Omega \to \mathbb{R}^n$. The vector neighborhood filter can be written as

$$\mathrm{NF}_{h,\rho}u(x) = \frac{1}{C(x)} \int_{B_\rho(x)} u(y) e^{-\frac{\|u(y)-u(x)\|^2}{h^2}} dy,$$

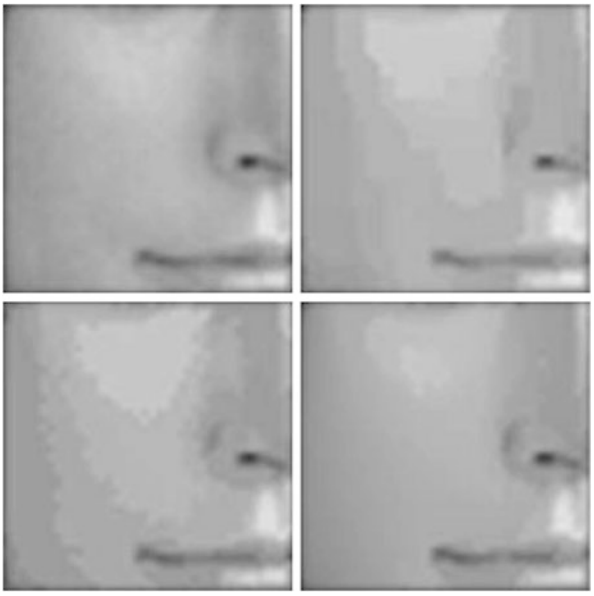


Fig. 22 Comparison experiment. *Top left*: original image. *Top right*: Perona-Malik filtered image. *Bottom left*: filtered image by the neighborhood filter. *Bottom right*: filtered image by the linear regression neighborhood filter. The neighborhood filter experiments are performed by iterating the discrete version of definitions (1) and (5). Both the neighborhood filter and its linear regression correction have been applied with the same value of h and ρ. The displayed images have been attained within the same number of iterations. The Perona-Malik equation is implemented by the explicit difference scheme proposed in the original paper. The Perona-Malik model and the neighborhood filter create artificial contours and flat zones. This effect is almost completely avoided by the linear regression neighborhood filter

where $||u(\mathbf{y}) - u(\mathbf{x})||^2$ is now the Euclidean vector norm and each component function u_i is filtered with the same weight distribution. The linear regression correction is defined as in the scalar case, and each component is locally approximated by a plane with the same weight distribution.

In order to compute the asymptotic expansion of the linear regression filter, we must fix a coordinate system for $\mathbb{R}^2$. In the scalar case, we used the reference system given by the gradient of the image at $\mathbf{x}$ and its orthogonal direction. In addition, this reference allows us to relate the obtained diffusion to the evolution of the level lines of the image and the mean curvature motion. Now, we cannot use the same reference and we need to define a new one. By analogy with the scalar case, we choose the directions of minimum and maximum variation of the vector function.

Definition 1. We define the normal direction η and the tangent direction ξ as the vectors that respectively maximize and minimize the following variation:

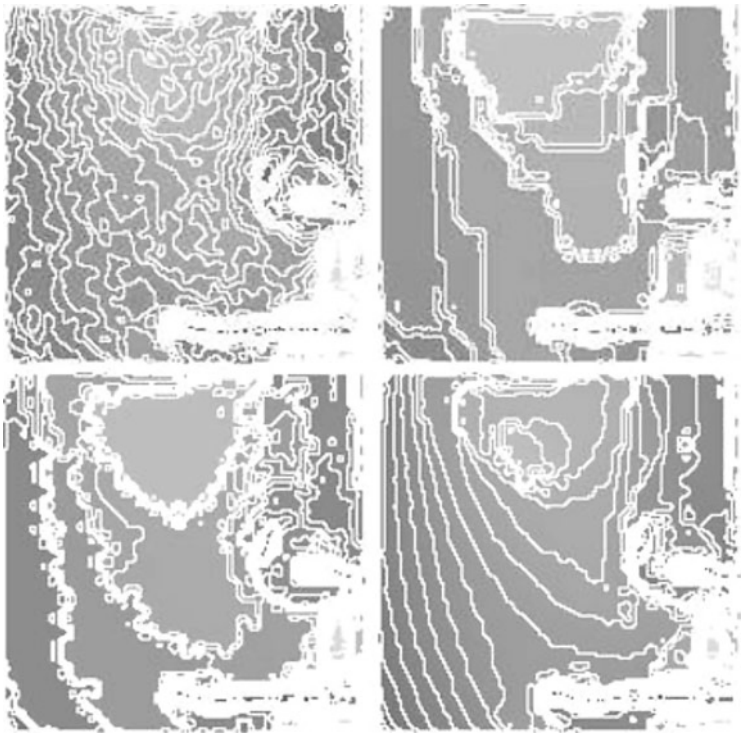


Fig. 23 Level lines of the images in Fig. 22. By the Perona-Malik filter and the neighborhood filter, the level lines tend to group, creating flat zones. The regression correction filters the level lines by a curvature motion without creating any flat zone

$$\sum_{i=1}^{n} \left\| \frac{\partial u_i}{\partial v}(x) \right\|^2$$

under the constraint $||v|| = 1$.

It is easily seen that this constrained optimization leads to the computation of the eigenvectors of the matrix

$$A = \begin{pmatrix} \left\| \frac{\partial u}{\partial x} \right\|^2 & \left\langle \frac{\partial u}{\partial x}, \frac{\partial u}{\partial y} \right\rangle \\ \left\langle \frac{\partial u}{\partial x}, \frac{\partial u}{\partial y} \right\rangle & \left\| \frac{\partial u}{\partial y} \right\|^2 \end{pmatrix},$$

where $\frac{\partial u}{\partial x} = \left(\frac{\partial u_1}{\partial x}, \ldots, \frac{\partial u_n}{\partial x} \right)$ and $\frac{\partial u}{\partial y} = \left(\frac{\partial u_1}{\partial y}, \ldots, \frac{\partial u_n}{\partial y} \right)$. The two positive eigenvalues of A, λ_+ and λ_-, are the maximum and the minimum of the vector norm associated to A and the maximum and the minimum variations, as defined in Definition 1. The corresponding eigenvectors are orthogonal leading to the above-defined normal and tangent directions. This orthonormal system was first proposed for vector-valued

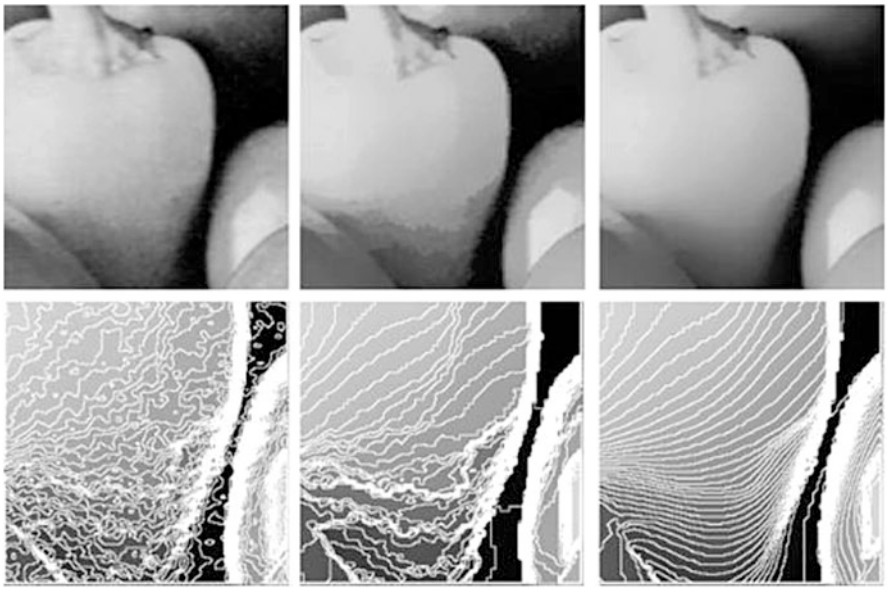


Fig. 24 Comparison of the neighborhood filter and the linear regression correction. *Top left*: original image. *Top middle*: filtered image by the neighborhood filter. *Top right*: filtered image by the regression neighborhood filter. *Bottom*: level lines of a part of the images on the above line. Both neighborhood filters have been performed with the same filtering parameters and the same number of iterations. The linear regression neighborhood algorithm has filtered the image while preserving the main boundaries as the original neighborhood filter. No enhancing has been applied by the linear correction avoiding the shock effect. The level lines of the neighborhood filter tend to group and create large flat zones. In addition, these level lines oscillate, while those of the linear regression algorithm have been correctly filtered

image analysis in [25]. Many PDE equations have been proposed for color image filtering using this system. We note the coherence-enhancing diffusion [72], the Beltrami flow [46], and an extension of the mean curvature motion [65].

Theorem 6. *Suppose* $u \in C^2(\Omega, \mathbb{R}^n)$, *and let* $\rho, h, \alpha > 0$ *such that* $\rho, h \to 0$ *and* $h = O(\rho^\alpha)$. *Let* $\tilde{f}$ *be the continuous function defined as* $\tilde{f}(0) = \frac{1}{6}$,

$$\tilde{f}(t) = \frac{1}{4t^2}\left(1 - \frac{2te^{-t^2}}{E(t)}\right),$$

for $t \neq 0$, *where* $E(t) = 2\int_0^t e^{-s^2} ds$. *Then, for* $\mathbf{x} \in \Omega$,

1. *If* $\alpha < 1$,

$$\mathrm{LNF}_{h,\rho}u(x) - u(x) \simeq \frac{\Delta u(x)}{6}\rho^2.$$

2. *If* $\alpha = 1$,

$$\mathrm{LNF}_{h,\rho}u(x) - u(x) \simeq \left[\tilde{f}\left(\frac{\rho}{h}\left\|\frac{\partial u}{\partial \xi}(x)\right\|\right) D^2u(\xi,\xi)(x) + \tilde{f}\left(\frac{\rho}{h}\left\|\frac{\partial u}{\partial \eta}(x)\right\|\right) D^2u(\eta,\eta)\right]$$

where $\Delta u(\mathbf{x}) = (\Delta u_i(\mathbf{x}))_{1\le i\le n}$ *and* $D^2u(v,v)(\mathbf{x}) = D^2u_i(v,v)(\mathbf{x})_{1\le i\le n}$ *for* $v \in \{\eta,\xi\}$.

Interpretation

When h is much larger than ρ, the linear regression neighborhood filter is equivalent to the heat equation applied independently to each component. When h and ρ have the same order, the subjacent PDE acts as an evolution equation with two terms. The first term is proportional to the second derivative of u in the tangent direction ξ. The second term is proportional to the second derivative of u in the normal direction η. The magnitude of each diffusion term depends on the variation in the respective direction, $\lambda_- = ||\frac{\partial u}{\partial \xi}(x)||$ and $\lambda_+ = ||\frac{\partial u}{\partial \eta}(x)||$. The weighting function $\tilde{f}$ is positive and decreases to zero (see Fig. 21). We can distinguish the following cases depending on the values of λ_+ and λ_-.

- If $\lambda_+ \simeq \lambda_- \simeq 0$, then there are very few variations of the vector image u around $\mathbf{x}$. In this case, the linear regression neighborhood filter behaves like a heat equation with maximum diffusion coefficient $\tilde{f}(0)$.
- If $\lambda_+ \gg \lambda_-$, then there are strong variations of u around $\mathbf{x}$ and the point may be located on an edge. In this case, the magnitude $\tilde{f}\left(\frac{\rho}{h}\lambda_+\right)$ tends to zero and there is no diffusion in the direction of maximal variation. If $\lambda_- \gg 0$, then $\mathbf{x}$ may be placed on an edge with different orientations depending on each component and the magnitude of the filtering in both directions tends to zero, so that the image is hardly altered. If $\lambda_- \simeq 0$, then the edges have similar orientations in all the components and the image is filtered by a directional Laplacian in the direction of minimal variation.
- If $\lambda_+ \simeq \lambda_- \gg 0$, then we may be located on a saddle point, and in this case the image is hardly modified. When dealing with multivalued images, one can think of the complementarity of the different channels leading to the perception of a corner.

In the scalar case, the theorem gives back the result studied in the previous sections. The normal and tangent directions are, respectively, the gradient direction and the level line direction. In this case, $\frac{\partial u}{\partial \xi}(x) = 0$ and $\frac{\partial u}{\partial \eta}(x) = |Du(x)|$, and we get back to

$$\mathrm{LNF}_{h,\rho}u(x) - u(x) \simeq \left[\frac{1}{6}D^2u(\xi,\xi)(x) + \tilde{f}\left(\frac{\rho}{h}|Du(x)|\right) D^2u(\eta,\eta)(x)\right]\rho^2$$

4 Variational and Linear Diffusion

The relationship of neighborhood filters with classic local PDEs has been discussed in the previous section. Yet, the main interest has shifted to defining *nonlocal PDEs*. The extension of the neighborhood filter and the NL-means method to define nonlocal image-adapted differential operators and nonlocal variational methods starts with [47], which proposes to perform denoising and deblurring by nonlocal functionals.

The general goal of this development is actually to give a variational to all neighborhood filters and to give a nonlocal form to the total variation as well. More precisely, the neighborhood filters derive from the functional

$$J(u) = \int_{\Omega\times\Omega} g\left(\frac{|u(x)-u(y)|^2}{h^2}\right) w\left(|x-y|\right) dxdy,$$

where g and w have a Gaussian decay. In the same line, a functional yields a (variational) interpretation to NL-means:

$$\mathrm{JNL}(u) = \int_{\Omega\times\Omega} \left(1 - e^{-\frac{G_\sigma * |u(x-.)-u(y-.)|^2(0)}{h^2}}\right) w\left(|x-y|\right) dxdy.$$

In a similar variational framework, Gilboa et al. [36] consider the general kind of quadratic nonlocal functional

$$J(u) := \int_{\Omega\times\Omega} (u(x)-u(y))^2 w\left(x-y\right) dxdy, \tag{6}$$

where $w(\mathbf{x},\mathbf{y})$ is any fixed weight distribution, which in most applications writes as the neighborhood or NL-means weight distribution. The resolution of the graph heat equation or the variational minimization (6) is given by

$$u_{n+1}(x) = \frac{1}{C(x)} \int_\Omega u_n(y) w(x,y) dy,$$

where $C(\mathbf{x}) = \int_\Omega w(\mathbf{x},\mathbf{y}) d\mathbf{y}$ is a normalizing factor. The freedom of having a totally decoupled weight distribution makes this formulation a linear and powerful tool for image processing. In fact, this formulation rewrites as the Dirichlet integral of the following nonlocal gradient: $\nabla_w u(x,y) = (u(\mathbf{x}) - u(\mathbf{y}))w(x,y)$. The whole process relates to a graph Laplacian where each pixel is considered as the node of a weighted graph, and the weights of the edge between two pixels x and y, respectively, are decreasing functions of the distances of patches around x and y, $w(x,y)$. Then a graph Laplacian can be calculated on this graph, seen as the sampling of a manifold, and the linear diffusion can be interpreted as the heat equation on the set of blocks endowed with these weights. The eigenvalues and

eigenvectors of such a Laplacian can be computed and used for designing spectral algorithms as Wiener and thresholding methods (see [69] and [59]).

The nonlocal term (6) has shown to be very useful as a regularization term for many image-processing tasks. The nonlocal differential operators permit to define a total variation or a Dirichlet integral. Several articles on deblurring have followed this variational line: [36, 44, 53] for image segmentation; [8] for fluorescence microscopy; [81], again, for nonlocal deconvolution; and [50] for deconvolution and tomographic reconstruction. In [33], a paper dedicated to another notoriously ill-posed problem, the super-resolution, the nonlocal variational principle, is viewed as "an emerging powerful family of regularization techniques," and the paper "proposes to use the example-based approach as a new regularising principle in ill-posed image-processing problems such as image super-resolution from several low resolution photographs." For all these methods, the weight distribution is computed in the first iteration and is maintained during the whole iteration process.

In this section, we will concentrate on the last nonlocal functional as a linear diffusion process and therefore the associated graph to the image as a heat equation; that is, no fidelity term will be added to the functional.

Linear Diffusion: Seed Growing

In [37,39], a novel method was proposed for performing multi-label, semiautomated medical image segmentation. The Grady segmentation method is a linearized sigma filter applied to propagate seed regions.

Given a small number of pixels with user-defined labels which are called seeds, this method computes the probability that a random walker starting at each unlabeled pixel will first reach one of the pre-labeled pixels. By assigning each pixel to the label for which the greatest probability is calculated, a high-quality image segmentation can be obtained.

With each unlabeled pixel, a K-tuple vector is assigned that represents the probability that a random walker starting from this unlabeled pixel first reaches each of the K seed points. A final segmentation may be derived from these K-tuples by selecting for each pixel the most probable seed destination for a random walker. By biasing the random walker to avoid crossing sharp intensity gradients, a quality segmentation is obtained that respects object boundaries (including weak boundaries). The image (or volume) is treated as a graph with a fixed number of vertices and edges. Each edge is assigned real-valued weight corresponding to the likelihood that a random walker will cross that edge (e.g., a weight of zero means that the walker may not move along that edge). By a classical result the probability that a random walker first reaches a seed point exactly equals the solution to the heat equation [9] with boundary Dirichlet conditions at the locations of the seed points, the seed point in question being fixed to unity, while the other seeds are set to zero.

This idea was not quite new. Region competition segmentation is an old concept [82]. One can also refer to an algorithm developed for machine learning by Zhu et al. [83], which also finds clusters based upon harmonic functions, using boundary

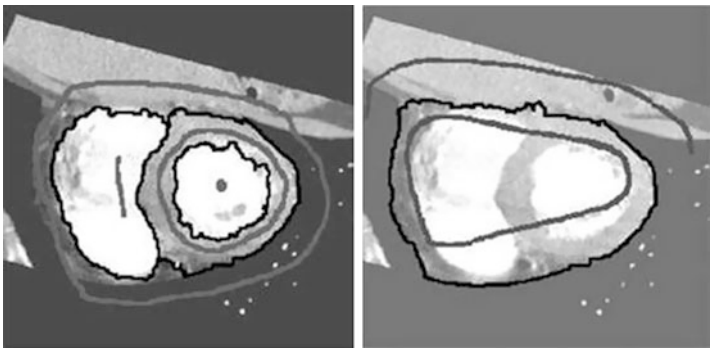


Fig. 25 (Taken from [38].) The Grady segmentation method is a linearized sigma filter applied to propagate seed regions. The *gray curves* are user-defined seed regions. A diffusion with sigma filter weights computed on the original image u_0 is applied until a steady state is attained. A threshold gives the black curves separating the regions of initial seeds

conditions set by a few seed points. Ref. [67] also involves weights in the image considered as a graph and takes seed points. The method is also directly related to the recent image-coloring method of Sapiro et al. by diffusion from seeds [78] (see also [65]).

Thus, the Grady segmentation method is a linearized sigma filter applied to propagate seed regions. Figure 25 taken from [38] illustrates the process on a two-chamber view of a cardiac image. The gray curves are user-defined seed regions roughly denoting the ventricles in the image. In that case, one of the seed regions is put to 1 and the other to 0. A diffusion with sigma filter weights computed on the original image u_0 is applied until a steady state is attained. This gives at each pixel $\mathbf{y}$ a value $p_1(\mathbf{y})$ between 0 and 1, which is interpreted as the probability for $\mathbf{y}$ to belong to the region of the first seed. In this binary case, a single threshold at 0.5 gives the black curves separating the regions of both seeds. Like the active contour method, this method is highly dependent on the initial seeds. It is, however, much less sensitive to noise than the snakes method [16] and permits to initialize fairly far from the desired contours. We will see that by the histogram concentration phenomenon, one can get similar or better results without any initialization.

The very same process as illustrated allows to diffuse initial chromatic information on an initial gray image as we exposed in the introduction. Figure 26 illustrates this application and compares the obtained solution by using the NL-means and the neighborhood filter.

Linear Diffusion: Histogram Concentration

The segmentation process can be accomplished by iterating the neighborhood filter and computing the weight distribution in the initial image, as displayed in Fig. 27. The top image shows one slice of a 3D CT image with interest area surrounded

Fig. 26 *Left* and from *top* to *bottom*: initial chromatic data on the gray image, linear diffused seeds by using neighborhood filter weights on the gray image, and the same for the NL-means weights. *Right*: details of left-hand images. The neighborhood filter weights are not robust since just a single point from different objects can be easily confused and iteration may lead to an incorrect colorization

by a parallelepiped. The next row shows several slices of this area of interest. It can be appreciated, first, that the background of arteries has a lot of oscillating clutter and, second, that the gray level value in arteries varies a lot, thus making an automatic threshold problematic. The best way actually to convince oneself that even in this small area a direct threshold would not do the job is to refer to the histograms of Fig. 29. The first histogram that is Gaussian-like and poorly concentrated corresponds to the background. The background mode decreases slowly. On the far right part of the histogram, one can see a small pick corresponding to very white arteries. The fixing of an accurate threshold in the slowly decreasing background mode is problematic. The top right histogram shows what happens after the application of a median iterative filtering (the mean curvature motion). The histogram does not concentrate at all. The bottom left histogram is obtained after

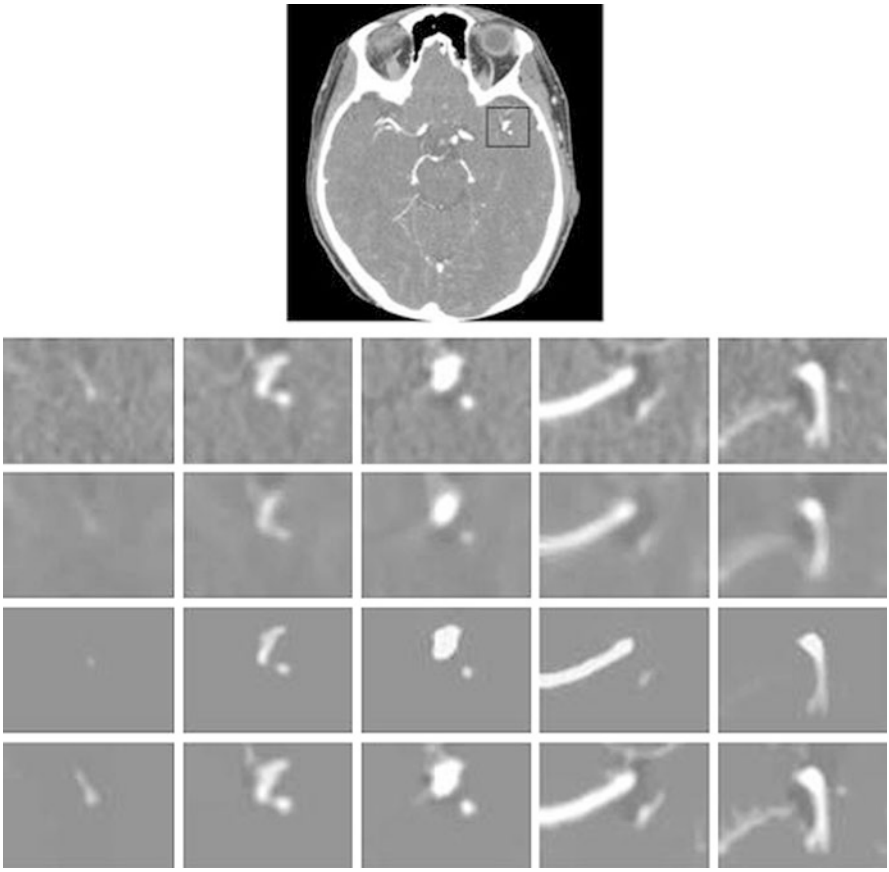


Fig. 27 Comparative behavior of discussed methods in 3D. Application to a 3D angiography CT image of the head where blood vessels should be segmented. *Top*: one slice image of the CT volume data with marked interested area. *Middle*: display of interest area for several slices of the 3D image. *Second row*: filtered slices by using median filter. *Third row*: sigma filter. *Fourth row*: 3D nonlocal heat equation. *Bottom*: filtered slices by using the linear method with 3D NL-means weights. The whole sequence has been treated as a 3D image with a weight support of $(5 \times 5 \times 3)$ and a comparison window of $3 \times 3 \times 3$. The background is flattened and blood vessels are enhanced. Thus, a better segmentation is possible by a simple threshold, as justified by Fig. 29

applying the linearized neighborhood filter. The bottom right histogram is the one obtained by the linearized NL-means described in the same section. In both cases, one observes that the background mode of the histogram is strongly concentrated on a few gray level values. An automatic threshold is easily fixed by taking the first local minimum after the main histogram peak. This histogram concentration is very similar to the obtained by the mean-shift approach [20] where the neighborhood filter is nonlinearly iterated. In that case, the authors show that clusters tend to its mean, yielding piecewise constant image.

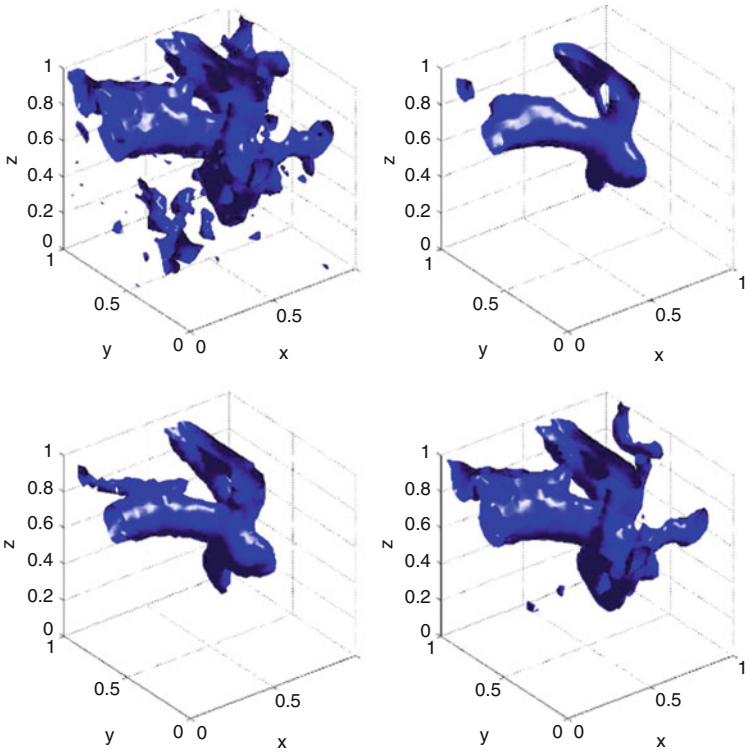


Fig. 28 From *top* to *bottom* and *left* to *right*: original iso-surface of the 3D image, same iso-surface filtered by iterative median filter, by linear sigma filter, and by linear NL-means. The iso-surface extracted from the original image presents many irregularities due to noise. The median filter makes them disappear, but makes important parts disappear and some vessels disconnect or fuse. Linear NL-means keeps most vessels and maintains the topology

The histogram concentration phenomenon is actually visible in the comparative evolution of some slices under the various considered filters, as shown in Fig. 27. The first row shows these slices picked in the interest area. The topology killing effect of the median filter (mean curvature motion)is as follows: small arteries tend to vanish and larger ones shrink and become circular as shown in the third slice showing an artery section. The third row is dedicated to the linear sigma filter, which corresponds to Grady's method applied directly to the image instead of using seeds. It is quite apparent that well-contrasted objects are well maintained and the contrast augmented, in agreement with the consistency of this recursive filter with the Perona-Malik equation. However, the less contrasted objects tend to vanish because, on them, the evolution becomes similar to an isotropic heat equation. The fourth row is the result of applying the 3D nonlocal linear heat equation, where the Laplacian coefficients are computed from the original image. The whole sequence has been treated as a 3D image with a weight support of $(7 \times 7 \times 3)$ and

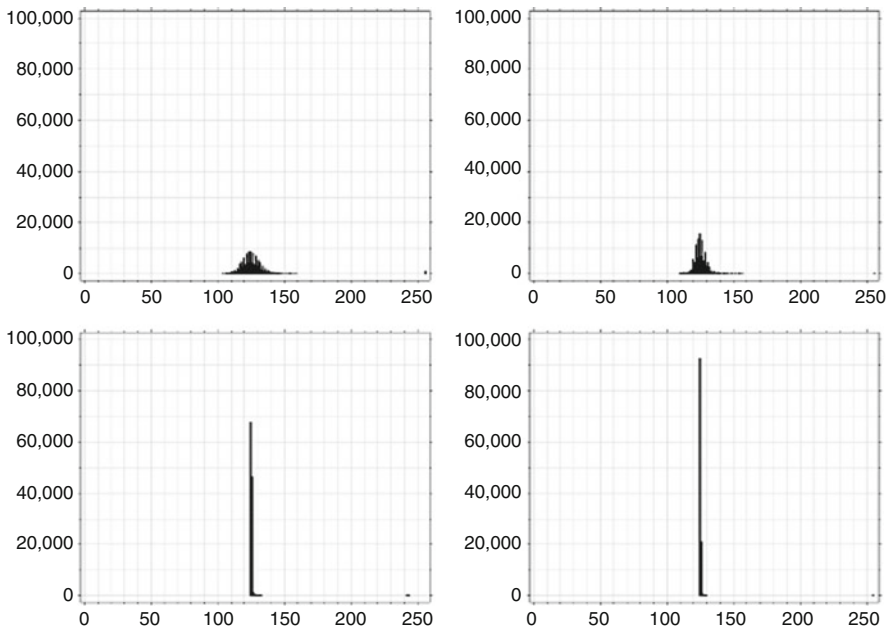


Fig. 29 Gray level histogram of 3D areas of interest. *Top left*: original 3D image before. *Top right*: after median filtering. *Bottom left*: after proposed method with sigma filter weights. *Bottom right*: proposed method with NL-means weights. The background is now represented by a few gray level values when the volume is filtered by the proposed method. A threshold can therefore be more easily and automatically applied

a comparison window of $3 \times 3 \times 3$. Clearly the background is flattened and blood vessels are enhanced on this background. A threshold just above the homogeneously made background level should give back arteries, and this indeed occurs. Thus, in that case, the 3D visualization of objects with complex topology like the cerebral arteries can be achieved by an automatic threshold as illustrated in Fig. 28. The exact segmentation of the artery is a more difficult problem. Even if the histogram is concentrated, a different choice of the visualization threshold can produce slightly different surfaces.

5 Conclusion

This chapter has introduced neighborhood filters and reviewed the impact that they have had in many image-processing problems during these last years.

The neighborhood filters have been analyzed under three different frameworks; a denoising filter, a local smoothing filter and in a variational formulation.

As a denoising algorithm, the denoising filter has motivated state of the art algorithms as its generalization the NL-means algorithm. As a local smoothing filter,

its asymptotic analysis leads to the well known Perona-Malik equation. And its variational formulation is effectively used for segmentation and diffusion purposes, for example, in medical image analysis.

Neighborhood filters remain as one of the main contributions in image processing of the last years and still influences current and future research.

Cross-References

▸ Neighborhood Filters and the Recovery of 3D Information
▸ Numerical Methods and Applications in Total Variation Image Restoration
▸ Total Variation in Imaging

References

1. Andreu, F., Ballester, C., Caselles, V., Mazon, J.M.: Minimizing total variation flow. Comptes Rendus de l'Academie des Sciences Series I Mathematics **331**(11), 867–872 (2000)
2. Arias, P., Caselles, V., Sappiro, G.: A variational framework for non-local image inpainting. In: Proceedings of the EMMCVPR, Bonn. Springer, Heidelberg (2009)
3. Attneave, F.: Some informational aspects of visual perception. Psychol. Rev. **61**(3), 183–193 (1954)
4. Aubert, G., Kornprobst, P.: Mathematical Problems in Image Processing: Partial Differential Equations and the Calculus of Variations. Springer, New York (2006)
5. Bae, S., Paris, S., Durand, F.: Two-scale tone management for photographic look. ACM Trans. Graph. (TOG) **25**(3), 645 (2006)
6. Barash, D.: A fundamental relationship between bilateral filtering, adaptive smoothing, and the nonlinear diffusion equation. IEEE Trans. Pattern Anal. Mach. Intell. **24**, 844–847 (2002)
7. Bennett, E.P., Mason, J.L., McMillan, L.: Multispectral bilateral video fusion. IEEE Trans. Image Process. **16**(5), 1185 (2007)
8. Boulanger, J., Sibarita, J.B., Kervrann, C., Bouthemy, P.: Nonparametric regression for patch-based fluorescence microscopy image sequence denoising. In: 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro, 2008 (ISBI 2008), Paris,pp. 748–751, 2008
9. Boykov, Y., Jolly, M.P.: Interactive graph cuts for optimal boundary and region segmentation of objects in ND images. Int. Conf. Comput. Vis. **1**, 105–112 (2001)
10. Brailean, J.C., Kleihorst, R.P., Efstratiadis, S., Katsaggelos, A.K., Lagendijk, R.L.: Noise reduction filters for dynamic image sequences: a review. Proc. IEEE **83**(9), 1272–1292 (1995)
11. Buades, A., Coll, B., Lisani, J., Sbert, C.: Conditional image diffusion. Inverse Probl. Imaging **1**(4):593 (2007)
12. Buades, A., Coll, B., Morel, J.M.: A review of image denoising algorithms, with a new one. Multiscale Model Simul. **4**(2), 490–530 (2005)
13. Buades, A., Coll, B., Morel, J.M.: Neighborhood filters and PDE's. Numer. Math. **105**(1), 1–34 (2006)
14. Buades, A., Coll, B., Morel, J.M.: Nonlocal image and movie denoising. Int. J. Comput. Vision. **76**(2), 123–139 (2008)
15. Buades, A., Coll, B., Morel, J.M., Sbert, C.: Self-similarity driven color demosaicking. IEEE Trans. Image Process. **18**(6), 1192–1202 (2009)
16. Caselles, V., Kimmel, R., Sapiro, G.: Geodesic active contours. Int. J. Comput. Vis. **22**(1), 61–79 (1997)

17. Choudhury, P., Tumblin, J.: The trilateral filter for high contrast images and meshes. In: ACM SIGGRAPH 2005 Courses, Los Angeles, p. 5. ACM (2005)
18. Cleveland, W.S.: Robust locally weighted regression and smoothing scatterplots. J. Am. Stat. Assoc. **74**(368), 829–836 (1979)
19. Coifman, R.R., Donoho, D.L.: Translation-Invariant De-Noising. Lecture Notes in Statistics, pp. 125–125. Springer, New York (1995)
20. Comaniciu, D., Meer, P.: Mean shift: a robust approach toward feature space analysis. IEEE Trans. Pattern Anal. Mach. Intell. **24**(5), 603–619 (2002)
21. Criminisi, A., Pérez, P., Toyama, K.: Region filling and object removal by exemplar-based image inpainting. IEEE Trans. Image Process. **13**(9), 1200–1212 (2004)
22. Danielyan, A., Foi, A., Katkovnik, V., Egiazarian, K.: Image and video super-resolution via spatially adaptive block-matching filtering. In: Proceedings of International Workshop on Local and Non-local Approximation in Image Processing, Lausanne (2008)
23. De Bonet, J.S.: Multiresolution sampling procedure for analysis and synthesis of texture images. In: Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques, Los Angeles, p. 368. ACM Press/Addison-Wesley (1997)
24. Delon, J., Desolneux, A.: Flicker stabilization in image sequences (2009). hal.archives-ouvertes.fr
25. Di Zenzo, S.: A note on the gradient of a multi-image. Comput. Vis. Graph. **33**(1), 116–125 (1986)
26. Dong, B., Ye, J., Osher, S., Dinov, I.: Level set based nonlocal surface restoration. Multiscale Model. Simul. **7**, 589 (2008)
27. Donoho, D.L.: De-noising by soft-thresholding. IEEE Trans. Inf. Theory **41**(3), 613–627 (1995)
28. Durand, F., Dorsey, J.: Fast bilateral filtering for the display of highdynamic-range images. In: Proceedings of the 29th annual conference on Computer graphics and interactive techniques, ACM New York, pp. 257–266 (2002)
29. Ebrahimi, M., Vrscay, E.R.: Solving the Inverse Problem of Image Zooming Using "Self-Examples". In: Kamel, M., Campilho, A. (eds.) Image Analysis and Recognition. Lecture Notes in Computer Science, vol. 4633, p. 117. Springer, Berlin/Heidelberg (2007)
30. Ebrahimi, M., Vrscay, E.R.: Multi-frame super-resolution with no explicit motion estimation. In: Proceedings of the 2008 International Conference on Image Processing, Computer Vision, and Pattern Recognition, Las Vegas, 2008
31. Efros, A.A., Leung, T.K.: Texture synthesis by non-parametric sampling. In: International Conference on Computer Vision, Corful, vol. 2, pp. 1033–1038, 1999
32. Eisemann, E., Durand, F.: Flash photography enhancement via intrinsic relighting. ACM Trans. Graph. (TOG) **23**(3), 673–678 (2004)
33. Elad, M., Datsenko, D.: Example-based regularization deployed to superresolution reconstruction of a single image. Comput. J. **50**, 1–16 (2007)
34. Elmoataz, A., Lezoray, O., Bougleux, S., Ta, V.T.: Unifying local and nonlocal processing with partial difference operators on weighted graphs. In: International Workshop on Local and Non-local Approximation in Image Processing, Lausanne (2008)
35. Fleishman, S., Drori, I., Cohen-Or, D.: Bilateral mesh denoising. ACM Trans. Graph. (TOG) **22**(3), 950–953 (2003)
36. Gilboa, G., Osher, S.: Nonlocal linear image regularization and supervised segmentation. Multiscale Model. Simul. **6**(2), 595–630 (2007)
37. Grady, L.: Random walks for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **28**(11), 1 (2006)
38. Grady, L., Funka-Lea, G.: Multi-label image segmentation for medical applications based on graph-theoretic electrical potentials. In: Computer Vision and Mathematical Methods in Medical and Biomedical Image Analysis (ECCV), pp. 230–245 (2004)
39. Grady, L.J.: Space-variant computer vision: a graph-theoretic approach. PhD thesis, Boston University (2004)

40. Guichard, F., Morel, J.M., Ryan, R.: Contrast invariant image analysis and PDEs. Book in preparation
41. Harten, A., Engquist, B., Osher, S., Chakravarthy, S.R.: Uniformly high order accurate essentially non-oscillatory schemes, III. J. Comput. Phys. **71**(2), 231–303 (1987)
42. Huhle, B., Schairer, T., Jenke, P., Straßer, W.: Robust non-local denoising of colored depth data. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Anchorage, pp. 1–7 (2008)
43. Jones, T.R., Durand, F., Desbrun, M.: Non-iterative, feature-preserving mesh smoothing. ACM Trans. Graph. **22**(3), 943–949 (2003)
44. Jung, M., Vese, L.A.: Nonlocal variational image deblurring models in the presence of Gaussian or impulse noise. In: Tai, X.-C., Mørken, K., Lysaker, M., Lie, K.-A. (eds.) Scale Space and Variational Methods in Computer Vision. Lecture Notes in Computer Science, vol. 5567. Springer, Berlin/Heidelberg (2009)
45. Kimia, B.B., Tannenbaum, A., Zucker, S.W.: On the evolution of curves via a function of curvature, I: the classical case. J. Math. Anal. Appl. **163**(2), 438–458 (1992)
46. Kimmel, R., Malladi, R., Sochen, N.: Images as embedded maps and minimal surfaces: movies, color, texture, and volumetric medical images. Int. J. Comput. Vis. **39**(2), 111–129 (2000)
47. Kindermann, S., Osher, S., Jones, P.W.: Deblurring and denoising of images by nonlocal functionals. Multiscale Model. Simul. **4**(4), 1091–1115 (2005)
48. Lee, J.S.: Digital image smoothing and the sigma filter. Comput. Vis. Graph. **24**(2), 255–269 (1983)
49. Lezoray, O., Ta, V.T., Elmoataz, A.: Nonlocal graph regularization for image colorization. In: International Conference on Pattern Recognition, Tampa (2008)
50. Lou, Y., Zhang, X., Osher, S., Bertozzi, A.: Image recovery via nonlocal operators. J. Sci. Comput. **42**(2), 185–197 (2010)
51. Mairal, J., Elad, M., Sapiro, G., et al.: Sparse representation for color image restoration. IEEE Trans. Image Process. **17**(1), 53 (2008)
52. Masnou, S.: Filtrage et désocclusion d'images par méthodes d'ensembles de niveau. PhD thesis, Ceremade, Universit e Paris-Dauphine (1998)
53. Mignotte, M.: A non-local regularization strategy for image deconvolution. Pattern Recognit. Lett. **29**(16), 2206–2212 (2008)
54. Osher, S., Rudin, L.I.: Feature-oriented image enhancement using shock filters. SIAM J Numer Anal **27**(4), 919–940 (1990)
55. Ozkan, M.K., Sezan, M.I., Tekalp, A.M.: Adaptive motion-compensated filtering of noisy image sequences. IEEE Trans. Circuits Syst. Video Technol. **3**(4), 277–290 (1993)
56. Peng, H., Rao, R., Messinger, D.W.: Spatio-spectral bilateral filters for hyperspectral imaging. In: SPIE Defense and Security Symposium. International Society for Optics and Photonics (2008)
57. Perona, P., Malik, J.: Scale-space and edge detection using anisotropic diffusion. IEEE Trans. PAMI **12**(7), 629–639 (1990)
58. Petschnigg, G., Szeliski, R., Agrawala, M., Cohen, M., Hoppe, H., Toyama, K.: Digital photography with flash and no-flash image pairs. ACM Trans. Graph. (TOG) **23**(2), 664–672 (2004)
59. Peyré, G.: Manifold models for signals and images. Comput. Vis. Image Underst. **113**(2),249–260 (2009)
60. Peyré, G.: Sparse modeling of textures. J. Math. Imaging Vis. **34**(1), 17–31 (2009)
61. Polzehl, J., Spokoiny, V.: Varying coefficient regression modeling by adaptive weights smoothing. WIAS preprint no. 818 (2003)
62. Protter, M., Elad, M., Takeda, H., Milanfar, P.: Generalizing the nonlocal-means to super-resolution reconstruction. IEEE Trans. Image Process. **18**(1), 35–51 (2009)
63. Ramanath, R., Snyder, W.E.: Adaptive demosaicking. J. Electron. Imaging **12**, 633 (2003)
64. Rudin, L., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. Phys. D **60**(1–4), 259–268 (1992)

65. Sapiro, G., Ringach, D.L.: Anisotropic diffusion of multivalued images with applications to color filtering. IEEE Trans. Image Process. **5**(11), 1582–1586 (1996)
66. Sethian, J.A.: Curvature and the evolution of fronts. Commun. Math. Phys. **101**(4), 487–499 (1985)
67. Shi, J., Malik, J.: Normalized cuts and image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **22**(8), 888–905 (2000)
68. Smith, S.M., Brady, J.M.: SUSAN: a new approach to low level image processing. Int. J. Comput. Vis. **23**(1), 45–78 (1997)
69. Szlam, A.D., Maggioni, M., Coifman, R.R.: A general framework for adaptive regularization based on diffusion processes on graphs. Yale technical report (2006)
70. Tomasi, C., Manduchi, R.: Bilateral filtering for gray and color images. In: Proceedings of the Sixth International Conference on Computer Vision, Bombay, pp. 839–846, 1998
71. van den Boomgaard, R., van de Weijer, J.: On the equivalence of localmode finding, robust estimation and mean-shift analysis as used in early vision tasks. In: International conference on pattern recognition, Quebec, vol. 16, Citeseer, pp. 927–930, 2002
72. Weickert, J.: Anisotropic Diffusion in Image Processing. B.G. Teubner, Stuttgart (1998). Citeseer
73. Winnemoller, H., Olsen, S.C., Gooch, B.: Real-time video abstraction. ACM Trans. Graph. (TOG) **25**(3), 1226 (2006)
74. Wong, A., Orchard, J.: A nonlocal-means approach to exemplar-based inpainting. In: 15th IEEE International Conference on Image Processing, San Diego, 2008, pp. 2600–2603
75. Yaroslavsky, L.P.: Digital Picture Processing. Springer Secaucus. Springer, Berlin/New York (1985)
76. Yaroslavsky, L.P.: Local adaptive image restoration and enhancement with the use of DFT and DCT in a running window. In: Unser, M.A., Aldroubi, A., Laine, A.F. (eds.) Wavelet Applications in Signal and Image Processing IV. Proceedings of SPIE, vol. 2825, p. 2. SPIE International Society for Optics and Photonics, Bellingham (1996)
77. Yaroslavsky, L.P., Egiazarian, K.O., Astola, J.T.: Transform domain image restoration methods: review, comparison, and interpretation. In: Dougherty, E.R., Astola, J.T. (eds.) Nonlinear Image Processing and Pattern Analysis XII. Proceedings of SPIE, vol. 4304, p. 155. SPIE International Society for Optics and Photonics, Bellingham (2001)
78. Yatziv, L., Sapiro, G.: Fast image and video colorization using chrominance blending. IEEE Trans. Image Process. **15**(5), 1120–1129 (2006)
79. Yoshizawa, S., Belyaev, A., Seidel, H.P.: Smoothing by example: Mesh denoising by averaging with similarity-based weights. In: IEEE International Conference on Shape Modeling and Applications, Matsushima, pp. 38–44, 2006
80. Zhang, D., Wang, Z.: Image information restoration based on longrange correlation. IEEE Trans. Circuits Syst. Video Technol. **12**(5), 331–341 (2002)
81. Zhang, X., Burger, M., Bresson, X., Osher, S.: Bregmanized nonlocal regularization for deconvolution and sparse reconstruction. UCLA CAM report 09-03 (2009)
82. Zhu, S.C., Yuille, A.: Region competition: unifying snakes, region growing, and bayes/MDL for multiband image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **18**(9), 884–900 (1996)
83. Zhu, X., Lafferty, J., Ghahramani, Z.: Semi-supervised learning: from Gaussian fields to gaussian processes. School of Computer Science, Carnegie Mellon University (2003)

Neighborhood Filters and the Recovery of 3D Information

Julie Digne, Mariella Dimiccoli, Neus Sabater, and Philippe Salembier

Contents

1 Introduction 1646
2 Bilateral Filters Processing Meshed 3D Surfaces 1647
Glossary and Notation 1647
Bilateral Filter Definitions 1648
Trilateral Filters 1652
Similarity Filters 1653
Summary of 3D Mesh Bilateral Filter Definitions 1655
Comparison of Bilateral Filter and Mean Curvature Motion Filter on Artificial Shapes 1656
Comparison of the Bilateral Filter and the Mean Curvature Motion Filter on Real Shapes 1657
3 Depth-Oriented Applications 1658
Bilateral Filter for Improving the Depth Map Provided by Stereo Matching Algorithms 1660
Bilateral Filter for Enhancing the Resolution of Low-Quality Range Images 1663
Bilateral Filter for the Global Integration of Local Depth Information 1666

J. Digne (✉)
LIRIS, Centre National de la Recherche Scientifique (CNRS), Lyon, France
École normale supérieure de Cachan, CMLA, Cachan, France
e-mail: julie.digne@liris.cnrs.fr

M. Dimiccoli
Image Processing Group, Pompeu Fabra University (UPF), Barcelona, Spain
e-mail: polmariella@hotmail.com

N. Sabater
CMLA, École normale supérieure de Cachan, Cachan, France
e-mail: neus.sabater@cmla.ens-cachan.fr

P. Salembier
Department of Signal and Communication, Universitat Politechnica de Catalunya, Barcelona, Spain
e-mail: philippe@gps.tsc.upc.edu; philippe.salembier@upc.edu

© Springer Science+Business Media New York 2015

O. Scherzer (ed.), *Handbook of Mathematical Methods in Imaging*,
DOI 10.1007/978-1-4939-0790-8_27

4 Conclusion 1671
Cross-References 1671
References 1671

Abstract

Following their success in image processing (see Chapter ▶ Local Smoothing Neighborhood Filters), neighborhood filters have been extended to 3D surface processing. This adaptation is not straightforward. It has led to several variants for surfaces depending on whether the surface is defined as a mesh, or as a raw data point set. The image gray level in the bilateral similarity measure is replaced by a geometric information such as the normal or the curvature. The first section of this chapter reviews the variants of 3D mesh bilateral filters and compares them to the simplest possible isotropic filter, the mean curvature motion.

In a second part, this chapter reviews applications of the bilateral filter to a data composed of a sparse depth map (or of depth cues) and of the image on which they have been computed. Such sparse depth cues can be obtained by stereovision or by psychophysical techniques. The underlying assumption to these applications is that pixels with similar intensity around a region are likely to have similar depths. Therefore, when diffusing depth information with a bilateral filter based on locality and color similarity, the discontinuities in depth are assured to be consistent with the color discontinuities, which is generally a desirable property. In the reviewed applications, this ends up with the reconstruction of a dense perceptual depth map from the joint data of an image and of depth cues.

1 Introduction

The idea of processing a pixel relatively to its similar looking neighbors proved to be very powerful and was adapted to solve a huge variety of problems. Since its primary goal is to denoise data and since the same denoising problem appeared for 3-dimensional surfaces, the idea of a 3D bilateral filter was only natural. Nevertheless, we shall see that this extension is far from straightforward. Multiple adaptations have in fact been introduced; experimental results show that it is far better for denoising a shape while preserving edges than an isotropic filter (as one could expect).

The bilateral filter could be used not only to filter images but also to diffuse information across an image: in numerous applications some information (e.g., depth value) is given only at some point positions. The problem is then to extrapolate the information for all pixels in the image. This can be used to improve the quality of disparity maps obtained by stereoscopy or to diffuse depth cues in images.

In the present chapter the different applications will be reviewed and tested experimentally. Section 2 reviews bilateral filters applied to 3D data point sets, often organized in a triangulation (a mesh). It ends up with comparative simulations

illustrating the advantage of bilateral filters on isotropic filtering. Section 3 considers the various cases where, in an image, depth values or depth cues are available and shows that the bilateral filter used as a diffusion tool performs well in restoring a dense depth map.

A previous review by Paris et al. [33] discusses the bilateral filter and its implementation. It also provides an overview of numerous applications.

2 Bilateral Filters Processing Meshed 3D Surfaces

This section proceeds by first examining the various adaptations of bilateral filtering on meshes (triangulated 3D surfaces) and discussing their implementation, which can depend on the surface triangulation. Finally several comparative experiments on synthetic and real meshes will be performed. Since a common notation is needed for all methods, this section starts with a small glossary and notation summary to which the reader may refer.

Glossary and Notation

- $\mathcal{M}$: the mesh, namely, a set of triangles
- v: current mesh vertex to be denoised
- $\mathcal{N}(v)$: neighborhood of vertex v (this neighborhood excludes v).
- n_v, n_p, etc.: normals at vertex v or point p, etc., to the underlying surface
- $w_1(||p - v||)$, $w_2(< n_v, p - v >)$, etc.: 1D centered Gaussians with various variances, used as weighting functions applied to the distance of neighbors to the current vertex and to the distance along the normal direction at v.
- H_v, H_p, etc.: curvatures of the underlying surface at v, p, etc.
- f: triangle of a mesh
- a_f: area of triangle f
- c_f: barycenter of triangle f
- n_f: normal to triangle f
- Π_f: projection on the plane containing triangle f
- V: voxel containing points of the data set
- s', v', p', n'_v: processed versions of s, v, p, n_v, etc.
- $\|p - q\|$: Euclidean distance between points p and q

The *neighborhood filter or sigma filter* is attributed to Lee [26] in 1983 but goes back to Yaroslavsky and the Sovietic image processing theory (see the book summarizing these works [44]) in 2D image analysis. A recent variant by Tomasi and Manduchi names it bilateral filter [37]. The bilateral filter denoises a pixel by using a weighted mean of its similar neighbors gray levels. In the original article, the similarity measure was the difference of pixel gray levels, yielding for a pixel v of an image I with neighborhood $\mathcal{N}(v)$:

$$\hat{I}(v) = \frac{1}{C(v)} \sum_{p \in \mathcal{N}(v)} w_1(\|p - v\|) w_2(|I(v) - I(p)|) I(p)$$

where w_1 et w_2 are decreasing functions on $\mathbb{R}^+$ (e.g., Gaussian) and $C(v)$ is a normalizing coefficient: $C(v) = \sum_{p \in \mathcal{N}(v)} w_1(\|p - v\|) w_2(|I(v) - I(p)|)$. Thus $\hat{I}(v)$ is an average of pixel values for pixels that are similar in position but also in value, hence the "bilaterality."

Bilateral Filter Definitions

Filtering without losing the sharp features is as critical for surfaces as it is for images, and a first adaptation of the bilateral filter to surface meshes was proposed by Fleishman, Drori, and Cohen-Or in [15]. Consider a meshed surface $\mathcal{M}$ with known normals n_v at each vertex position v. Let $\mathcal{N}(v)$ be the one-ring neighborhood of v (i.e., the set of vertices sharing an edge with v). Then the filtered position of v writes $v' = v + \delta v \cdot n_v$, where

$$\delta v = \frac{1}{C(v)} \sum_{p \in \mathcal{N}(v)} w_1(\|p - v\|) w_2(< n_v, p - v >) < n_v, p - v > \tag{1}$$

where the weight normalization factor is $C(v) = \sum_{p \in \mathcal{N}(v)} w_1(\|p - v\|) w_2(< n_v, p - v >)$. In a nutshell, this means that the normal component of the vertex v is moved by a weighted average of the normal components of its neighboring points which are also close to the plane tangent to the surface at v. The distance to the tangent plane plays for meshes the role that was played for images by the distance between gray levels. If v belongs to a sharp edge, then the only points close to the tangent plane at v are the points on the edge. Thus, the edge sharpness will not be smoothed away. One of the drawbacks of the above filter is clearly the use of a mesh-dependent neighborhood. In case of a mesh with fixed length edges, using the one-ring neighborhood is the same as using a fixed size neighborhood. Yet in most cases mesh edges do not have the same length. The one-ring neighborhood is then very dependent on the mesh representation and not on the shape itself. This is easily fixed by defining an intrinsic Euclidean neighborhood.

Another adaptation of the 2D bilateral filter to surface meshes is introduced by Jones, Durand, and Desbrun in [20]. This approach considers the bilateral filtering problem as a robust estimation problem for the vertex position. A set of surface predictors are linked to the mesh $\mathcal{M}$: for each triangle f the position estimator Π_f projects a point to the plane defined by f. Let a_f be the surface area and c_f be the center of f. Then, for each vertex v, the denoised vertex is

$$v' = \frac{1}{C(v)} \sum_{f \in \mathcal{M}} \Pi_f(v) a_f w_1(\|c_f - v\|) w_2(\|\Pi_f(v) - v\|) \tag{2}$$

where $C(v) = \sum_{f \in \mathcal{M}} a_f w_1(\|c_f - v\|) w_2(\|\Pi_f(v) - v\|)$ is the weight normalizing factor and w_1 and w_2 are two Gaussians.

Thus, the weight $w_1(||c_f - v||)$ is small if the triangle f is close to v. This term is the classic locality-in-space term of the bilateral. Similarly, $w_2(\|\Pi_f(v) - v\|)$ measures how far point v is from its projection onto the plane of the triangle. This weight favors the triangles f whose plane is coherent with v.

Since the projection on the tangent planes operator Π_f depends on the normals to f, these normals must be robustly estimated. Normals being first-order derivatives, they are more subject to noise than vertex positions. Hence the method starts by denoising the normal field. To do so, the mesh is first smoothed using the same formula as above without the influence weight w_2 and with $\Pi_f(v) = c_f$, namely, an updated position:

$$v' = \frac{1}{C(v)} \sum_{f \in \mathcal{M}} c_f a_f w_1(\|c_f - v\|)$$

where $C(v) = \sum_{f \in \mathcal{M}} a_f w_1(\|c_f - v\|)$. The normal for each face in the denoised mesh is then computed and assigned to the corresponding face of the original noisy mesh. It is with this robust normal field that the bilateral filter of Eq. (2) is applied in a second step.

The idea of filtering normals instead of point positions is crucial in point rendering applications, as was pointed out by Jones, Durand, and Zwicker in [21]. Indeed, when rendering a point set, removing noise from normal is more important than removing noise from point position, since normal variations are in fact what is perceived by observers. More precisely the eye perceives a dot product of the illumination and the normal, which makes it very sensitive to noisy normal orientations. The bilateral filter of [20] is seen as a deformation F of the points: $v' = F(v)$. Then, the update of normal n_v can be obtained through the transposed inverse of the Jacobian $J(v)$ of $F(v)$:

$$n'_v = J^{-T}(v) n_v, \text{ where } J_i(v) = \frac{\partial F}{\partial v_i}(v)$$

where J_i is the ith column of J and v_i is the ith component of v. n_v must then be renormalized. The rendering of the point set with smoothed normal is better than without any smoothing.

In [38], Wang introduces a related bilateral approach which denoises feature-insensitive sampled meshes. Feature insensitive means that the mesh sampling is independent of the features of the underlying surface, e.g., uniform sampling. The algorithm proceeds as follows: it detects the shape geometry (namely, sharp regions), denoises the points, and finally optimizes the mesh by removing thin triangles. The bilateral filter is defined in a manner similar to [20], with the difference that only triangles inside a given neighborhood are used on this definition. Let v be a mesh vertex, $\mathcal{N}(v)$ be the set of triangles within a given range of v, and

n_f, a_f, c_f be, respectively, the normal, area, and center of a facet f (a triangle). Denote by $\Pi_f(v)$ the projection of v onto the plane of f, and then the denoised vertex is defined by

$$v' = \frac{1}{C(v)} \sum_{f \in \mathcal{N}(v)} \Pi_f(v) a_f w_1(\|c_f - v\|) w_2(\|\Pi_f(v) - v\|)$$

where $C(v) = \sum_{f \in \mathcal{N}(v)} a_f w_1(\|c_f - v\|) w_2(\|\Pi_f(v) - v\|)$ (weight normalizing factor).

The first step is to detect sharp regions. Several steps of bilateral filtering (as defined in [20]) are applied, and then a smoothness index is computed by measuring the infimum of angles between normals of faces adjacent to v. By thresholding this measurement, the sharp vertices are selected. Triangles whose three vertices are sharp and whose size does not increase during the bilateral iterations are marked as sharp. This detection done, points are restored to their original positions. Then the bilateral filtering formula is applied to sharp vertices only, and the geometry sharpness is encoded into a data collection containing normals, centers, and areas of filtered triangles. Points are then restored to their original position. Each sharp vertex is moved using the bilateral filtering over the neighboring stored data units, and thin vertices are removed from the mesh (these last two steps are iterated a certain number of times). Finally, a post-filtering step consists in applying one step of bilateral filtering on all non-sharp edges.

In [40] (Wang, Yuan, and Chen), a two-step denoising method combines the fuzzy C-means clustering method (see Dunn's article on fuzzy means [12]) with a bilateral filtering approach. Fuzzy C-means is a clustering technique that allows a piece of data to belong to two different clusters. Each point p gets a parameter $\mu_{p,k}$ which measures the degree of membership of p to a cluster k. Let m_p be the number of points in the spherical neighborhood of a point p. If $m_p <$ threshold, the point is deleted. Otherwise, a fuzzy C-means clustering center c_p is associated with p. The normal at point c_p is computed as the normal to the regression plane of the data set in a spherical neighborhood of p. Fleishman's bilateral filter [15] is used to filter c_i which yields the denoised point. This hybrid and complex method is doubly bilateral. Indeed, the previous C-means clustering selects an adapted neighborhood for each point and replaces it by an average which is by itself the result of a first bilateral filter in the wide sense of neighborhood filter. Indeed, the used neighborhood for each point depends on the point. The second part of the method therefore applies a second classical bilateral method to a cloud that has been filtered by a first bilateral filter.

The bilateral filtering idea was also used as a part of a surface reconstruction process. In [30], for example, Miropolsky and Fischer introduced a method for reducing position and sampling noise in point cloud data while reconstructing the surface. A 3D geometric bilateral filter method for edge-preserving and data reduction is introduced. Starting from a point cloud, the points are classified in an octree, whose leaf cells are called *voxels*. The voxel centers are filtered,

representative surface points are defined, and the mesh is finally reconstructed. A key point is that the denoising depends on the voxel decomposition. Indeed, the filter outputs a result for each voxel. For a voxel V, call v its centroid with normal n_v. Let w_1 and u_2 be two functions weighting, respectively, $\|p - v\|$, the distance between a point p position and the centroid location, and $\delta(p, v) = \langle n_p, n_v \rangle$, the scalar product of the normal at p and the normal at the centroid. Then the output of the filter for voxel V is

$$v' = \frac{1}{C(v)} \sum_{p \in V} w_1(\|p - v\|) u_2(\delta(p, v)) p$$

where $C(v) = \sum_{p \in V} w_1(\|p - v\|) u_2(\delta(p, v))$. Here w_1 is typically a Gaussian and u_2 is an increasing function on $[0, 1]$. But this filter proves unable to recover sharp edges, so a modification is introduced: prior to any filtering for each voxel V, points of V are projected onto a sphere centered at the centroid v. Each mapped point is given a normal $\tilde{n}_p$ which has direction $p - v$ and is normalized. The geometric filtering is reduced to

$$v' = \frac{1}{C(v)} \sum_{p \in V} u_2(\delta(\tilde{n}_p, n_v)) p \text{ with } C(v) = \sum_{p \in V} u_2(\delta(\tilde{n}_p, n_v)).$$

Although only the similarity of normals is taken into account in the above formula, the filter is bilateral because the average is localized in the voxel.

In [27], Liu et al. interpreted the bilateral filter as the association to each vertex v of a weighted average

$$v' = \frac{1}{C(v)} \sum_{p \in \mathcal{N}(v)} w_1(\|p - v\|) w_2(\|\Pi_p(v) - v\|) \Pi_p(v)$$

where $C(v) = \sum_{p \in \mathcal{N}(v)} w_1(\|p - v\|) w_2(\|\Pi_p(v) - v\|)$ (normalizing factor) and $\Pi_p(v)$ is a *predictor* which defines a "denoised position of v due to p," namely, the *projection of v on the plane passing by p and having the normal n_v*. For example, the bilateral predictor used in [15] is $\Pi_p(v) = v + ((p - v) \cdot n_v) n_v$, and in [20], the used predictor is $\Pi_p(v) = v + ((p - v) \cdot n_p) n_p$ which is the projection of v on the tangent plane passing by p. With this last predictor the corners are less smoothed out, yet there is a tangential drift due to the fact that the motion is not in the normal direction n_v but in an averaged direction of the n_p for $p \in \mathcal{N}(v)$. Therefore a new predictor is introduced:

$$\Pi_p(v) = v + \frac{(p - v) \cdot n_p}{n_v \cdot n_p} n_v.$$

This predictor tends to preserve better the edges than all other bilateral filters.

The question of choosing automatically the parameters for the bilateral filter was raised by Hou, Bai, and Wang in [19]. It was proposed to choose adaptive parameters. The adaptive bilateral normal smoothing process starts by searching for the set of triangles $(T_i)_i$ whose barycenters are within a given distance of a center triangle T. (But this keeps a distance parameter anyway.) Then the influence weight parameter σ_s is computed as the standard deviation of the distance between normals $\|n(T_i) - n(T)\|$. The spatial weight parameter is estimated using a minimum length descriptor criterion (for various scales). The estimated parameters are then used to get the smoothed normal. This result is finally used for rebuilding the mesh using the smoothed normals by Ohtake, Belyaev, and Seidel's method described in [31].

The bilateral filter has proved to be very efficient to denoise a mesh while preserving sharp features. The trilateral filter is then a natural extension which takes into account still more geometric information.

Trilateral Filters

Choudhury and Tumblin [6] propose an extension of the trilateral image filter to oriented meshes. It is a 2-pass filter: a first pass filters the normals and a second pass filters the vertex positions. Starting from an oriented mesh, a first pass denoises bilaterally the vertex normals using the following update:

$$n'_v = \frac{1}{C(n_v)} \sum_{p \in \mathcal{N}(v)} n_p w_1(\|p - v\|) w_2(\|n_p - n_v\|)$$

where $C(n_v) = \sum_{p \in \mathcal{N}(v)} w_1(\|p - v\|) w_2(\|n_p - n_v\|)$. Then, an adaptive neighborhood $\mathcal{N}(v)$ is found by iteratively adding faces near v until the normals n_f of face f differ too much from n'_v. A function F measuring the similarity between normals is built using a given threshold R:

$$F(v, f) = 1 \text{ if } \|n'_v - n_f\| < R; 0 \text{ otherwise.}$$

The trilateral filter for normals filters a difference between normals. Define $n_\Delta(f) = n_f - n'_v$. Then the trilaterally filtered normal n_v is

$$n''_v = n'_v + \frac{1}{C(v)} \sum_{f \in \mathcal{N}(v)} n_\Delta(f) w_1(\|c_f - v\|) w_2(n_\Delta(f)) F(v, f)$$

where $C(v) = \sum_{f \in \mathcal{N}(v)} w_1(\|c_f - v\|) w_2(n_\Delta(f)) F(v, f)$. Finally, the same trilateral filter can be applied to vertices. Call P_v the plane passing through v and orthogonal to n'_v. Call $\tilde{c}_f$ the projection of c_f onto P_v and $c_\Delta(f) = \|\tilde{c}_f - c_f\|$. Then the trilateral filter for vertices, using the trilaterally filtered normal n''_v, writes

$$v' = v + n''_v \frac{1}{C(v)} \sum_{p \in \mathcal{N}(v)} c_\Delta(f) w_1(\|\tilde{c}_f - v\|) w_2(n_\Delta(f)) F(v, f)$$

where $C(v) = \sum_{p \in \mathcal{N}(v)} w_1(\|\tilde{c}_f - v\|) w_2(c_\Delta(f)) F(v, f)$.

The results are similar to [20] though slightly better. They are comparable to the results of [15] since both methods use the distance to the tangent plane as a similarity between points.

Similarity Filters

In [41] Wang et al. proposed a trilateral filter with slightly different principles. A *geometric intensity* of each sampled point is first defined as depending on the neighborhood of the point

$$\delta(p) = \frac{1}{C(p)} \sum_{q \in \mathcal{N}(p)} w_{pq} < n_p, q - p >$$

with

$$w_{pq} = w_1(\|q - p\|) w_2(\| < n_p, q - p > \|) w_h(\|H_q - H_p\|)$$

and

$$C(p) = \sum_{q \in \mathcal{N}(p)} w_{pq}.$$

This type of filter is a trilateral filter, which means that it depends on three variables: distance between the point p and its neighbors q, distance along the normal n_p between the point p and its neighbors q, and the difference of their mean curvatures H_p and H_q.

At each point, a local grid is built on the local tangent plane (obtained by local covariance analysis), and at each point of this grid, the geometry intensity is defined by interpolation. Thus, neighborhoods of the same geometry are defined for each pair of distinct points, and the similarity can be computed as a decreasing function of the L^2 distance between these neighborhoods.

Since the goal is to denoise one point with similar points, the algorithm proposes to cluster the points into various classes by the mean shift algorithm. To denoise a point, only points of the same class are used. This gives a denoised geometry intensity $\delta'(p)$ and the final denoised position $p' = p + \delta'(p) n_p$.

More recently the NL-means (Buades, Coll, Morel [3]) method which proved very powerful in image denoising was adapted to meshes and point clouds by Yoshizawa, Belyaev, and Seidel [47]. Recall that for an image I, the NL-means filter computes a filtered value $J(x)$ of pixel x as

$$J(x) = \frac{1}{C(x)} \int_{\Omega} w(x, y) I(y) dy,$$

an adaptive average with weights

$$w(x, y) = \exp -\frac{1}{h^2} \int G_a(|t|) |I(x - t) - I(y - t)|^2 dt$$

and $C(x) = \int_{\Omega} w(x, y) dy$.

Here G_a is a Gaussian or a compactly supported function, so that it defines a patch. Thus, the denoised point is a mean of pixel values with weights measuring the local image similarity of patches around other pixels with the patch around the current pixel.

Consider now the adaptation to a mesh $\mathcal{M}$. Let $\Omega_\sigma(v) = \{y \in \mathcal{M} | \quad |v - y| \leq 2\sigma\}$. The smoothing is done by changing v at each step: $v^{n+1} = v^n + k(v^n) n_v^n$ with n_v the normal to $\mathcal{M}$ at v. Let S_y be the surface associated to vertex y. The following definitions are directly adapted from the image case (a continuous formalism is adopted here for clarity):

$$k(v) = \frac{1}{C(v)} \int_{\Omega_{\sigma_2}} w(v, y) I(y) dS_y$$

$$C(v) = \int_{\Omega_{\sigma_2}} w(v, y) dS_y$$

$$I(y) = < n_v, y - v >$$

$$w(v, y) = \exp -\frac{D(v, y)}{h^2}.$$

The problem is to define the similarity kernel D. Let σ_3 be the half radius of the neighborhood used to define the geometric similarity between two points, and σ_2 be the half radius of the domain where similar points are looked for, with $\sigma_3 < \sigma_2$. The local tangent plane at y is parameterized by t_1 and t_2. For all z of $\Omega_{\sigma_2}(y)$, the translation t is defined as $t = -(< t_1, z - y >, < t_2, z - y >) = -(u_z, v_z)$, where (u_z, v_z, w_z) are the coordinates of vertex z in the local coordinate system (y, t_1, t_2, n_y).

A local approximation $F_v(u, v)$ by radial basis functions (RBF) is built around each vertex v, and the similarity kernel finally yields

$$D(v, y) = \int_{\Omega_{\sigma_3}(y)} G_{\sigma_3}(|t|) |F_v(u_z, v_z) - I(y - t)|^2 dt$$

with $I(y - t) = < n_v, z - v >$ and G_{σ_3} a Gaussian kernel with variance σ_3.

Thus each vertex is compared with vertices in a limited domain around it, and the weighted mean over all these nodes yields the denoised position. This results in

a better feature preserving mesh denoising method, but at the cost of a considerably higher computation time.

To improve the computation time when denoising data using neighborhood filters, bilateral approximations were introduced by Paris and Durand, among others, in [32], where a signal processing interpretation of the 2D bilateral filter is given, yielding an efficient approximation. Another efficient method is the Gaussian k-d trees introduced by Adams et al. in [1]. The proposed method was designed to compute efficiently a class of n-dimensional filters which replace a pixel value by a linear combination of other pixel values. The basic idea is to consider those filters as nearest neighbors search in a higher-dimensional space, for example, (r, g, b, x, y) in case of a 2D color image and a bilateral filter. To accelerate this neighbor search, a Gaussian k-d tree is introduced. Consider the nonlocal means filter which has, in its naive implementation, a $O(n^2 f^2)$ complexity for n pixels and $f \times f$ patches. To apply Gaussian k-d tree, the position of a pixel is set to be the patch, and the value is set to be the color value of the pixel. A simple Principle Component Analysis (PCA) on patches helps to capture the dimensions that best describe the patches. The Gaussian k-d tree is also used to perform 3D NL-means on meshes or point clouds. To produce a meaningful value to describe geometry, the idea of spin images is used. At each point sample p, a regression plane is estimated, and the coordinates of the neighboring points in the local coordinate system are used to build a histogram of cylindrical coordinates around (p, n_p) (the spin image). This gives the position vector. The value of p is then set to be the difference $d = p' - p$ between p and the Laplacian filtered position p' expressed in the local coordinate system. This gives the input for building the Gaussian k-d tree yielding good results for mesh denoising.

Summary of 3D Mesh Bilateral Filter Definitions

The filters reviewed in this section are almost all defined for meshes. Yet, with very little effort almost all of them can be adapted to unstructured point clouds by simply redefining the neighborhoods as the set of points within a given distance from the center point (spherical neighborhood). Several classic variants of bilateral filters were examined, but their main principle is to perform an average of neighboring vertices pondered by the distance of these vertices to an estimated tangent plane of the current vertex. This distance takes the role played by the gray level in image bilateral filters. It can be implemented in several ways either by projecting the current vertex to the neighboring triangles or by projecting the neighboring vertices on the current triangle or by using an estimate of the normal at the current vertex which has been itself previously filtered. An interesting and simple possibility is to directly combine distance of vertices and of their normals or even distances of vertices, normals, and curvatures (but this requires a previous smoothing to get denoised normals and curvatures). Notice that position, normal, and curvature characterize the cloud shape in a larger neighborhood. Thus, at this point, the obvious generalization of bilateral filters is NL-means, which directly compares

point-wise the shape of the neighborhood of a vertex with the overall shape of the neighborhoods of others before performing an average of the most similar neighborhoods to deliver a filtered neighborhood.

Sticking to the simplicity of comparisons and to the essentials of bilateral filter, we shall be contented in the comparative section to illustrate the gains of the bilateral filter with respect to a (good) implementation of its unilateral counterpart, the mean curvature motion, performed by the projection of each vertex on a local regression plane. The remainder of this section is divided as follows: section "Comparison of Bilateral Filter and Mean Curvature Motion Filter on Artificial Shapes" presents experiments and comparisons on artificial shapes, and section "Comparison of the Bilateral Filter and the Mean Curvature Motion Filter on Real Shapes" presents results on some real shapes.

Comparison of Bilateral Filter and Mean Curvature Motion Filter on Artificial Shapes

In the following experiments, the denoising of the bilateral filter as introduced in [15] will be compared with the mean curvature motion (MCM). Recall that [15] defined the update of a point as

$$\delta v = \frac{1}{C(v)} \sum_{p \in \mathcal{N}(v)} w_1(\|p - v\|) w_2(< n_v, p - v >) < n_v, p - v >$$

with

$$C(v) = \sum_{p \in \mathcal{N}(v)} w_1(\|p - v\|) w_2(< n_v, p - v >)$$

(see first section, Eq. (1) for notations).

The mean curvature motion used here is the projection on the regression plane: a vertex v with normal n_v and spherical neighborhood $\mathcal{N}(v)$ is projected on the regression plane of $\mathcal{N}(v)$. In [9], Digne et al. showed that this operator was an approximation of the mean curvature motion:

$$\frac{\partial v}{\partial t} = H n_v.$$

The effects of bilateral denoising are first shown on some artificial shapes. A cube with sidelength 5 is created with added Gaussian noise with standard deviation 0.02 (Fig. 1). Figures 1 and 2 show all points of the 3D cloud seen from one side of the cube. Obviously, the edges seem to have some width due to the noise.

The experiments of Fig. 2a–c show the denoising power of the bilateral filter in terms of preserving edges and should be compared with the standard mean curvature motion filter (Fig. 2d–f). The comparison is particularly interesting in the

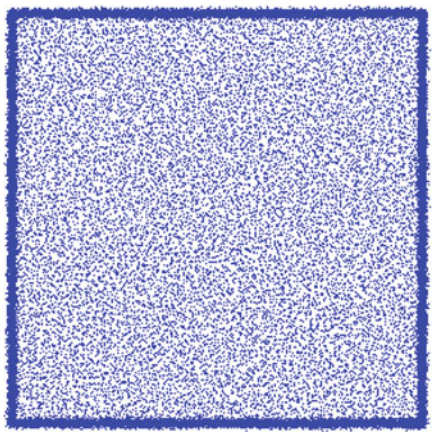


Fig. 1 A noisy cube with Gaussian noise

Table 1 Noise estimation for the sharp edge denoising

	Input	Iteration 1	Iteration 2	Iteration 5
RMSE (bilateral)	0.01	0.0031	0.0019	0.0035
RMSE (mcm)	0.01	0.0051	0.0085	0.0164

corner areas. The bilateral filter implies an anisotropic curvature motion leading to a diffusion only in smooth parts while preserving the sharp areas. Let us now see how those filters perform in case of a sharp edge. An estimation of the noise for each of the denoising methods is shown on Table 1. This estimation was obtained as follows: an edge was created by sampling two intersecting half-planes and adding Gaussian noise, the obtained edge was then denoised by bilaterally filtering an mean curvature motion. Finally, the root-mean-square error (RMSE) to the underlying model is computed. Table 1 tends to prove that mean curvature motion, although it smoothes well the noisy flat parts and also smoothes away the sharpness, whereas the bilateral filter tends to preserve the sharp edges better. With few iterations, the noisy parts are smoothed out decreasing the root-mean-square error. Then, when iterating the operator, the sharpness tends to be smoothed, increasing the RMSE again. This phenomenon is of course far quicker with the mean curvature motion since this filter does not preserve edges at all.

Comparison of the Bilateral Filter and the Mean Curvature Motion Filter on Real Shapes

This section starts with running some experiments on the Michelangelo's David point cloud. At each step an interpolating mesh was built for visualization.

On Fig. 3, denoising artifacts created by the bilateral filter can be seen. They appear as oscillations, for example, on David's cheek. These artifacts can be explained by the fact that the bilateral filter enhances structures. Added noise

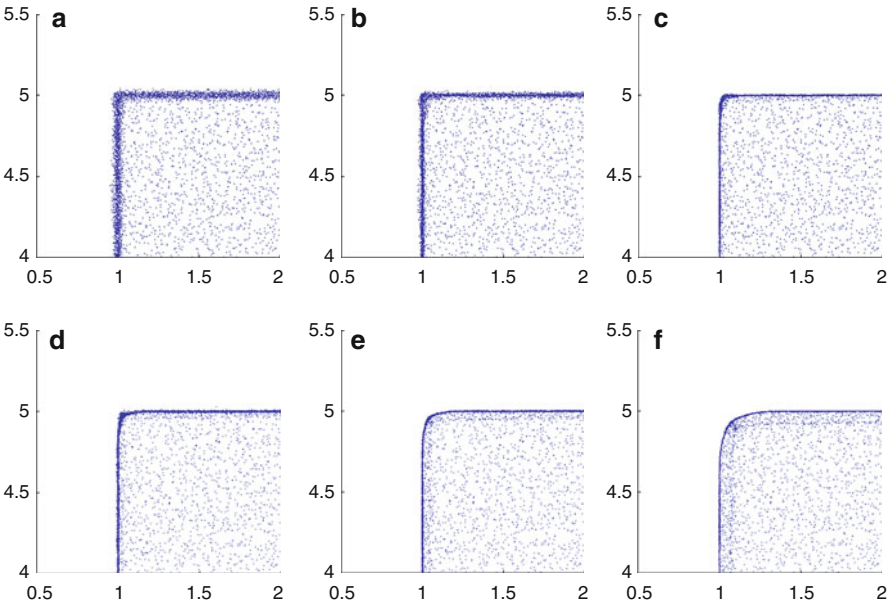


Fig. 2 Bilateral and MCM iterations on the cube corner. Notice how the sharpness is much better preserved by the bilateral filter than by the mean curvature equation. (**a**) 1 bilateral iteration. (**b**) 2 bilateral iterations. (**c**) 5 bilateral iterations. (**d**) 1 MCM iteration. (**e**) 2 MCM iterations. (**f**) 5 MCM iterations

structures can thus be randomly enhanced by the bilateral filter. Figure 4 shows that some noise remains after one iteration of bilateral denoising. The bilateral filter is therefore iterated with the same parameters. Then, obviously, the remaining noise disappears at the cost of some sharpness loss (see Fig. 5). Still, the bilateral filter preserves sharp features much better than the mean curvature motion (Fig. 6). This can also be seen on a noisy simple scan of a screw nut driver (Fig. 7) and on a fragment of the Stanford Digital Forma Urbis Romae Project (Fig. 8).

3 Depth-Oriented Applications

This section focuses on the applications of the bilateral filter and its generalized version to depth-oriented image processing tasks. The common idea to all these applications is to constrain the diffusion of depth information to the intensity similarity between pixels. The underlying assumption is that pixels with similar intensity around a region are likely to have similar depths. Therefore, when diffusing depth information based on intensity similarity, the discontinuities in depth are assured to be consistent with the color discontinuities. This is often a desirable property, as it was noticed by Gamble and Poggio [16] and Kellman and Shipley [22].

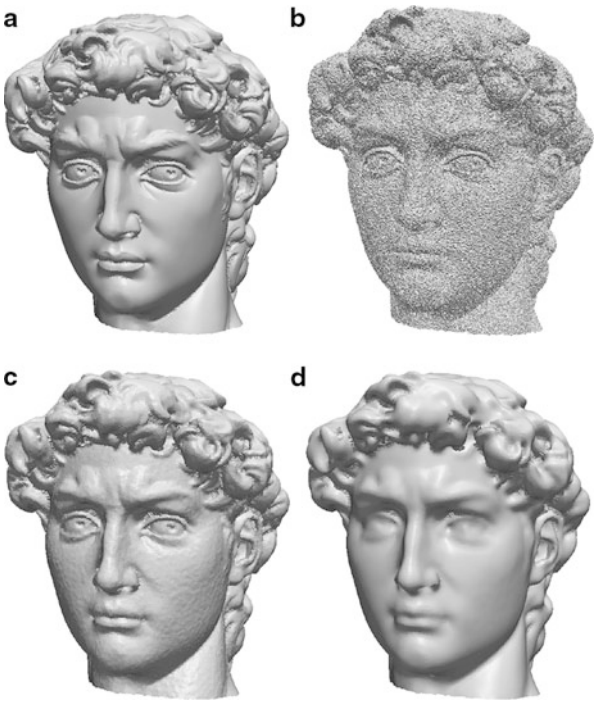


Fig. 3 Denoising of the David's face. (**a**) Original of the David. (**b**) Noisy David. (**c**) Bilateral denoising. (**d**) MCM

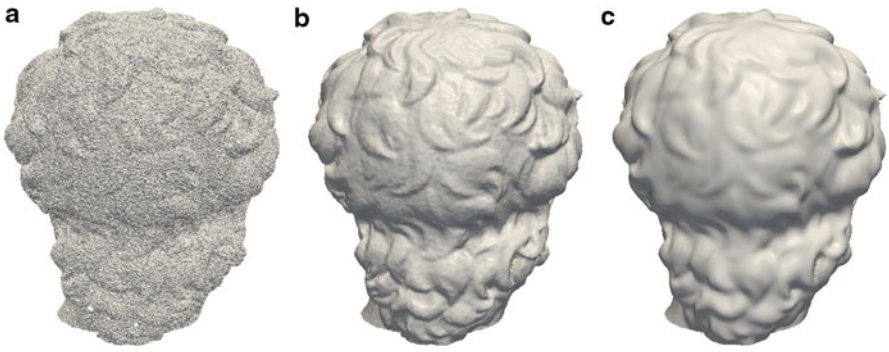


Fig. 4 Denoising of the David (back). (**a**) Initial noisy David. (**b**) Bilateral denoising. (**c**) MCM

The remainder of this section is organized as follows. Section "Bilateral Filter for Improving the Depth Map Provided by Stereo Matching Algorithms" reviews the applications of the bilateral filter to stereo matching algorithms, while section "Bilateral Filter for Enhancing the Resolution of Low-Quality Range

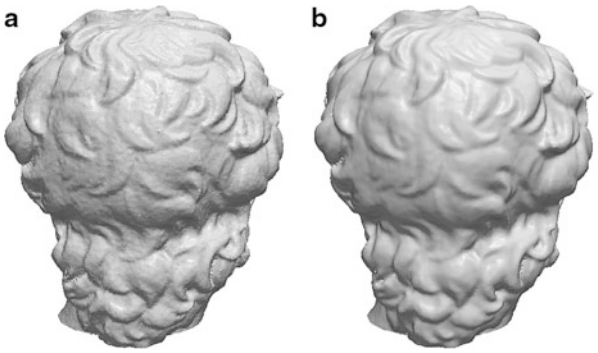


Fig. 5 Iterating the bilateral filter on the David (back). (**a**) Iteration 1. (**b**) Iteration 2

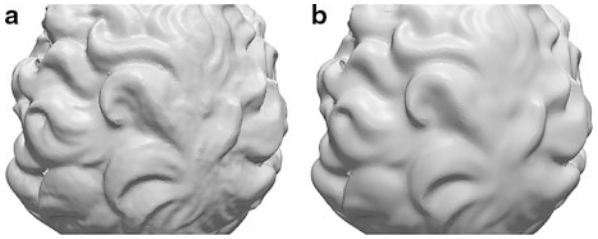


Fig. 6 Detail of the David. (**a**) Bilateral filtering. (**b**) MCM

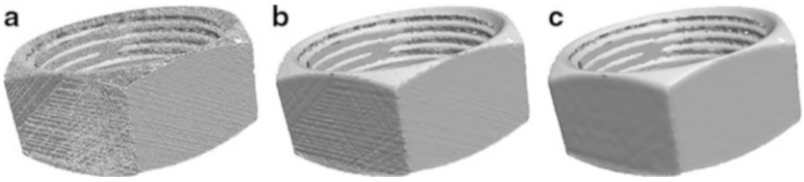


Fig. 7 Denoising of a screw nut driver scan. (**a**) Initial scan. (**b**) Bilateral denoising. (**c**) MCM

Images" describes an application to the resolution enhancement of range images. Section "Bilateral Filter for the Global Integration of Local Depth Information" reviews applications to the estimation of depth in single images.

Bilateral Filter for Improving the Depth Map Provided by Stereo Matching Algorithms

Stereo matching algorithms address the problem of recovering the depth map of a 3D scene from two images captured from different viewpoints. This is achieved by finding a set of points in one image which can be identified in the other one. In fact, the point-to-point correspondences allow to compute the relative disparities, which are directly related to the distance of the scene point to the image plane.

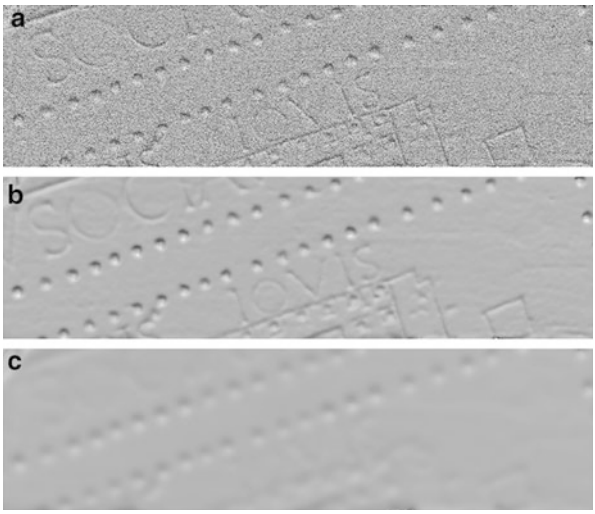


Fig. 8 Denoising of fragment "31u" of Stanford Digital Forma Urbis Romae; see Koller et al. [23] for an explanation of the data. (**a**) Initial fragment with added Gaussian noise. (**b**) Bilateral denoising. (**c**) MCM

The matching process is based on a similarity measure between pixels of both images. Due to the presence of noise and repetitive texture, these correspondences are extremely difficult to find without global reasoning. In addition, occluded and textureless regions are ambiguous. Indeed, local image matching is not enough to find reliable disparities in the whole image. Because of all these reasons, the matching process yields either low-accuracy dense disparity maps or high-accuracy sparse ones.

Improvements can be obtained through filtering or interpolation, by using median or morphological filters, for instance. However, their ability to do so is limited. Yin and Cooperstock have proposed [45] a post-processing step to improve dense depth maps produced by any stereo matching algorithm. The proposed method consists in applying an iterated bilateral filter, which diffuses the depth values. This diffusion relies on the original image gradient instead of the one of the depth images. This allows to incorporate edge information into the depth map, assuring discontinuities in depth to be consistent with intensity discontinuities.

The color-weighted correlation idea underlying the bilateral filter has been exploited by Yoon and Kweon [46] to reduce the ambiguity of the correspondence search problem. Classically, this problem has been addressed by area-based methods relying on the use of local support windows. In this approach, all pixels in a window are assumed to have similar depth in the scene and, therefore, similar disparities. Accordingly, pixels in homogeneous regions get assigned the disparities inferred from the disparities of neighboring pixels.

However, when the windows are located on depth discontinuities, the same disparity is assigned to pixels having different depths, resulting in a foreground-

fattening phenomenon. This phenomenon was studied by Delon and Rougé [8]. To obtain accurate results, an appropriate window should be selected for each pixel adaptively. This problem is addressed by Yoon and Kweon [46] by weighting the pixels in a given window taking into account their color similarity and geometric proximity to the reference pixel.

The similarity between two pixels is then measured using the support weights in both windows, taking into account the edge information into the disparity map. Experimental results show that the use of adaptive support weights produces accurate piecewise smooth disparity maps while preserving depth discontinuities.

The idea of exploiting the color-weighted correlation to reduce the ambiguity of the correspondence problem has been implemented in a parallel architecture, allowing its use in real-time applications and more complex stereo systems. See Yang et al. [42] and Wang et al. [39] papers which achieved a good rank in the Middlebury benchmark (Evaluation of the Middlebury stereo website http://vision.middlebury.edu/stereo/) proposed by Scharstein and Szeliski [36].

The bilateral filter averages the pixel colors, based on both their geometric closeness and their photometric similarity, preferring near values in space and color to distant ones. Ansar, Castano, and Matthies [2], Yoon and Kweon [46], and, more recently, Mattoccia, Giardino, and Gambin [28] have used the bilateral filter to weight the correlation windows before the stereo correspondence search. On the other hand, Gehrig and Franke [17] have applied the bilateral filter to obtain an improved and smoother disparity map.

The interpolation of disparity maps and in particular of digital elevation models (DEMs) has been considered in several recent works. Facciolo and Caselles [14] propose to interpolate unknown areas by constraining a diffusion anisotropic process to the geometry imposed by a reference image and coupling the process with a data fitting term which tries to adjust the reconstructed surface to the known data. More recently, Facciolo et al. [13] have proposed a new interpolation method which defines a geodesic neighborhood and fits an affine model at each point. The geodesic distance is used to find the set of points that are used to interpolate a piecewise affine model in the current sample. This interpolation is refined by merging the obtained affine patches with a Mumford-Shah-like algorithm. The a contrario methodology has been used in this merging procedure. In the urban context, Lafarge et al. [25] use a dictionary of complex building models to fit the disparity map. However, the applicability of such a method is less evident because of the initial delineation of buildings by a rectangle fitting.

We shall illustrate the bilateral interpolation process with experiments from Sabater's Ph.D. thesis [35] where the bilateral filter is used to interpolate a sparse disparity map. Let q be a point in the image I. Consider $L_q \subset I$ the subimage where the weight is learned. For each $p \in L_q$ the weight due to color similarity and proximity is computed.

Color similarity: the following color distance is considered

$$d_c(u_q, u_p) = \left((R_u(q) - R_u(p))^2 + (G_u(q) - G_u(p))^2 + (B_u(q) - B_u(p))^2\right)^{1/2},$$

where R_u, G_u, and B_u are the red, green, and blue channels of u. Then the weight corresponding to the color similarity between p and q is

$$w_c(p,q) = \exp\left(-\frac{d_c(u_q,u_p)^2}{h_1^2}\right).$$

Proximity: The Euclidean distance between the point positions in the image plane is used

$$d(q,p) = \left((q_1-p_1)^2+(q_2-p_2)^2\right)^{1/2},$$

where $p=(p_1,p_2)$ and $q=(q_1,q_2)$. Then the weight corresponding to proximity is

$$w_d(p,q) = \exp\left(-\frac{d(q,p)^2}{h_2^2}\right).$$

Therefore, the total associated weight between the two points q and p is

$$W(p,q) = \frac{1}{Z_q}\, w_c(p,q) w_d(p,q) = \frac{1}{Z_q} \exp\left(-\left(\frac{d_c(u_q,u_p)^2}{h_1^2}+\frac{d(q,p)^2}{h_2^2}\right)\right),$$

where Z_q is the normalizing factor $Z_q = \sum_{p\in L_q} w_c(p,q) w_d(p,q)$. The interpolated disparity map μ_I is computed via an iterative scheme

$$\mu_I(q,k) = \sum_{p\in L_q} W(p,q)\mu_I(p,k-1),$$

where k is the current iteration and the initialization $\mu_I(\cdot,0) = \mu(\cdot)$ is the sparse disparity to be interpolated.

Figures 9 and 10 show the interpolated Middlebury results (100 % density). The experiments demonstrate that, starting from a disparity map which is very sparse near image boundaries, the bilateral diffusion process can recover a reasonable depth map.

Bilateral Filter for Enhancing the Resolution of Low-Quality Range Images

Contrary to intensity images, each pixel of a range image expresses the distance between a known reference frame and a visible point in the scene. Range images are acquired by range sensors that, when acquired at video rate, are either very expensive or very limited in terms of resolution. To increase the resolution of

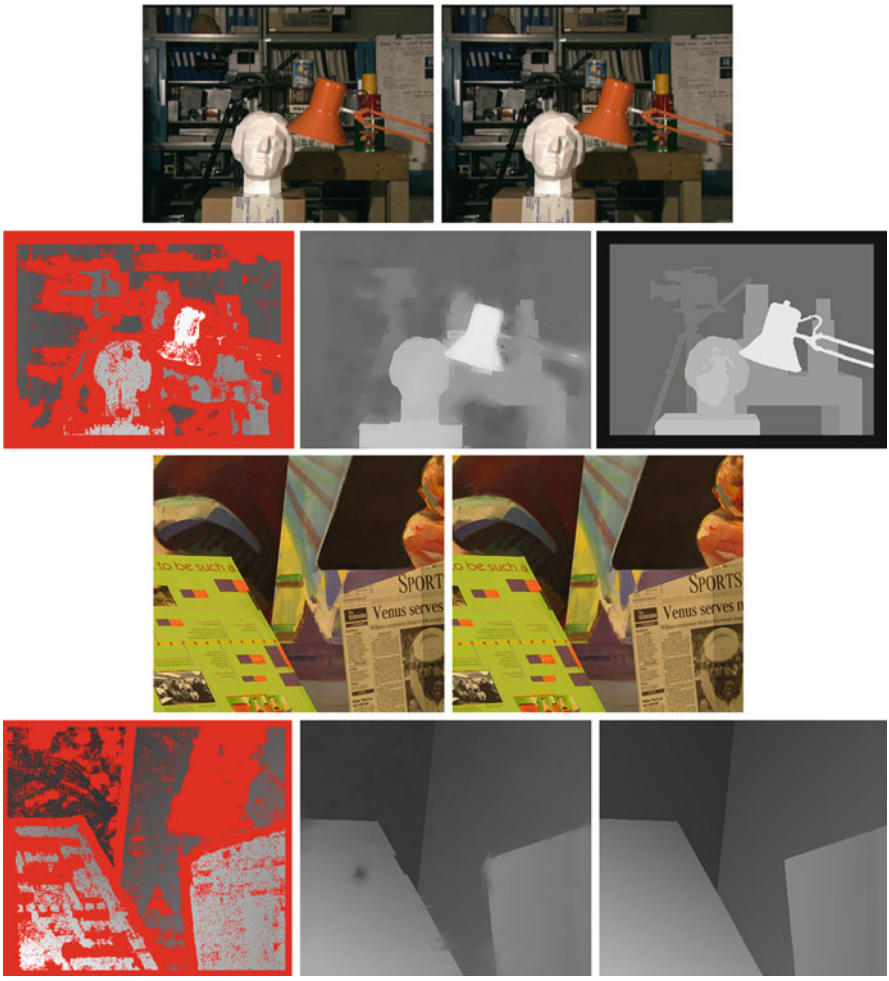


Fig. 9 Tsukuba and Venus results. For each couple of images: stereo pair of images, output of a sparse algorithm retaining only sure points, points (in *red*) are the rejected correspondences, interpolated version of these results and ground truth

low-quality range images acquired at video rate, Yang et al. [43] have proposed a post-processing step relying on an iterated bilateral filter. The filter diffuses the depth values of the low-quality range image, steering the diffusion by the color information provided by a registered high-quality camera image.

The input low-resolution range image is up-sampled to the camera image resolution. Then an iterative refinement process is applied. The up-sampled range image D_0 is used as the initial depth map to build an initial 3D cost volume c_0. The 3D cost volume $c_i(\mathbf{x}, \mathbf{y}, d)$ associated to the current depth map D_i at the ith iteration is given by

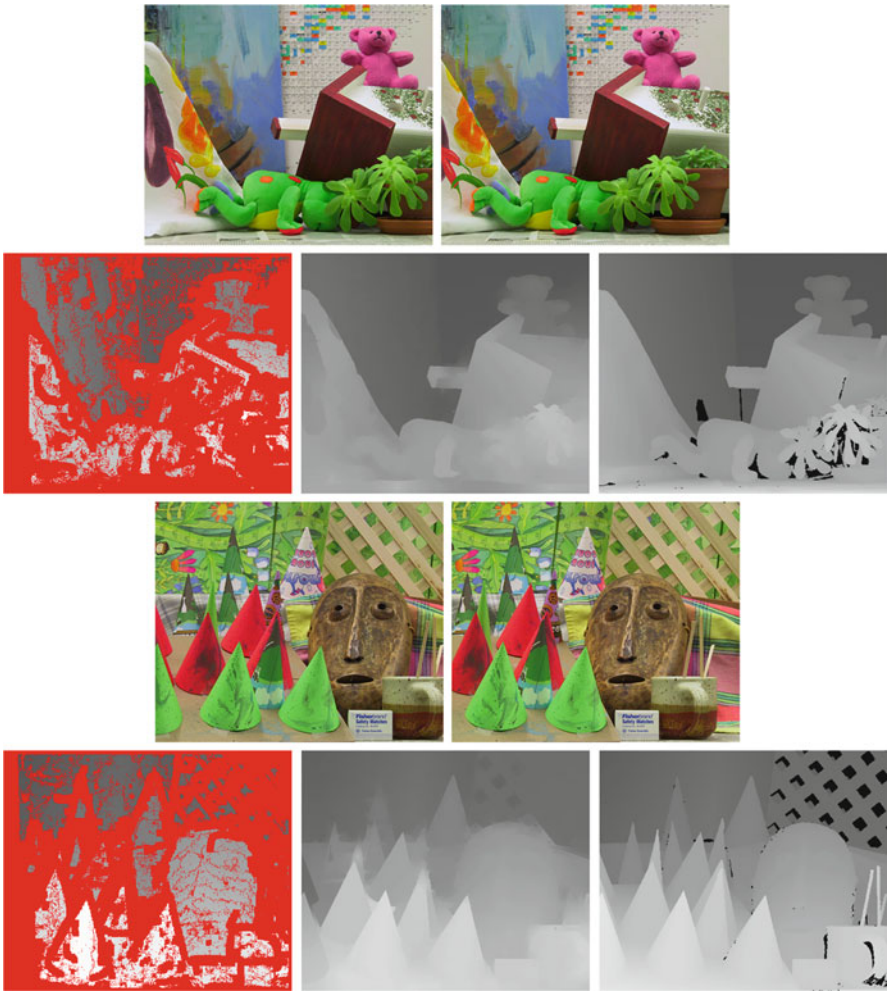


Fig. 10 Teddy and Cones results. For each couple of images: stereo pair of images, output of a sparse algorithm retaining only sure points, points (in *red*) are the rejected correspondences, interpolated version of these results and ground truth

$$c_i(\mathbf{x}, \mathbf{y}, d) = \min\left(\mu L, (d - D_i(\mathbf{x}, \mathbf{y}))^2\right) \tag{3}$$

where d is the depth candidate, L is the search range controlled by constant μ, and $D_i(\mathbf{x}, \mathbf{y})$ is the current depth estimate. To each depth candidate, d in the search range corresponds a single slice (disparity image) of the current cost volume. At each iteration, a bilateral filter is applied on each slice of the current cost volume c_i. This allows to smooth each slice image while preserving the edges. A new cost volume c_i^{BF} is therefore generated. Based on this new cost volume, a refined depth map D_{i+1} is obtained by selecting for each $(\mathbf{x}, \mathbf{y})$ the lowest cost candidate d.

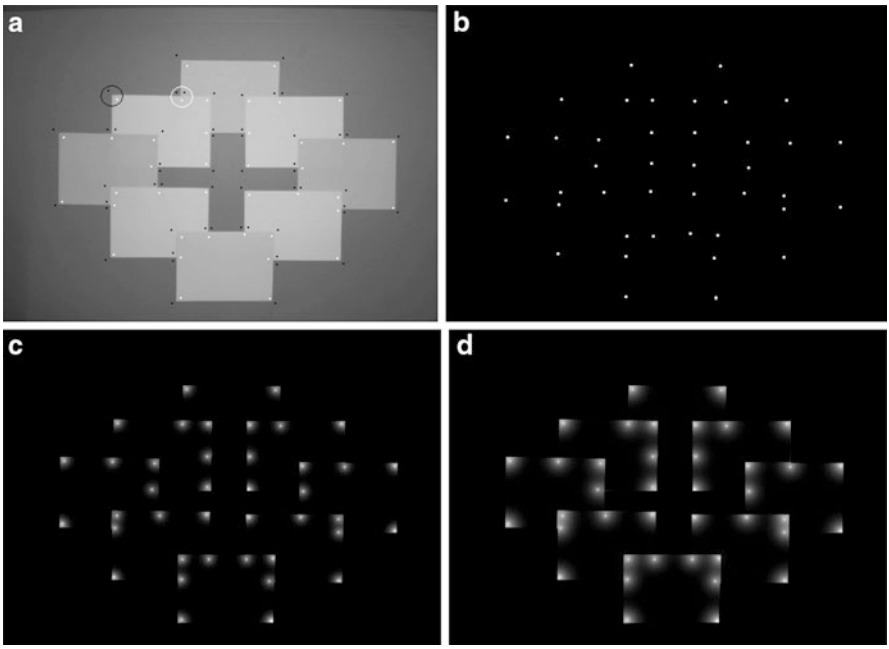


Fig. 11 Example of depth diffusion using Eq. (4). (**a**) *Gray*-level image, where BSPs and FSPs are marked in *white* and *black*, respectively. (**b**) Depth image, where points corresponding to FSPs are initialized with a positive value (marked in *white*) and the rest of the image with value zero. (**c**) and (**d**) Depth images after an increasing number of iterations of the DDF

Bilateral Filter for the Global Integration of Local Depth Information

Following the phenomenological approach of gestaltists [29], the perception of depth in single images results from the global integration of a set of monocular depth cues. However, all methods proposed in the computer vision literature to estimate depth in single *real* images rely on the use of prior experience about objects and their relationships to the environment. As a consequence, these methods generally rely on strong assumptions on the image structure [7,18], for instance, that the world is made of ground/horizontal planes and vertical walls, or assumptions on the image content [34] such as the prior knowledge of the class of objects being involved.

In contrast to the state of the art, Dimiccoli, Morel, and Salembier [11] proposed a general low-level approach for estimating depth in single real images. In this approach the global depth interpretation is directly inferred from a set of monocular depth cues, without relying on any previously learned contextual information nor on any strong assumption on the image structure. In particular the set of initial local depth hypothesis derived from different monocular depth cues is used to initialize and constrain a diffusion process. This diffusion is based on an iterated neighborhood filter.

With this strategy, the occlusion boundaries and the relative distances from the viewpoint of depicted objects are simultaneously recovered from local depth information, without the need of any explicit segmentation. Possible conflicting depth relationships are automatically solved by the diffusion process itself.

Once monocular depth cues are detected, each region involved in a depth relationship is marked by one or few points, called *source points* (see Fig. 11a). Source points marking the regions closer to the viewpoint are called Foreground Source Points (FSPs), whereas source points marking the regions more distant to the viewpoint are called Background Source Points (BSPs). In case of occlusion, three source points are marked (see white circle in Fig. 11a). A single FSP marks the region representing the occluding object, and two corresponding BSPs mark the partially occluded object and the background. In case of convexity, there is a single FSP and its corresponding BSP (see black circle in Fig. 11a). The depth image z is initialized by assigning a positive value Δ to FSPs and value 0 to BSPs. The rest of the image is initialized with value 0 (see Fig. 11b). The diffusion process is applied to the depth image z by using the gradient of the original image u rather than the one of the depth image. Doing so, the edge information is incorporated into the depth map, ensuring that depth discontinuities are consistent with gray-level (color) discontinuities.

The depth diffusion filter (DDF) proposed in [11] by Dimiccoli, Morel, and Salembier is

$$\mathrm{DDF}_{h,r} z(x) = \frac{1}{C(x)} \int_{S_r(x)} z(y) e^{\frac{-|u(x)-u(y)|^2}{h^2}} dy, \tag{4}$$

where $S_r(x)$ is a square of center x and side r, h is the filtering parameter which controls the decay of the exponential function, and

$$C(x) = \int_{S_r(x)} e^{\frac{-|u(x)-u(y)|^2}{h^2}} dy \tag{5}$$

is the normalization factor. In practice, the parameters are $r = 3$ and $h = 10$.

Equation (4) is applied iteratively until the stability is attained. In the discrete case, after each iteration, the values of FSPs and BSPs are updated. More precisely, if the difference between the values of an FSP and the corresponding BSP becomes smaller than Δ, then Δ is added to the value of the FSP. In the continuous case, the neighborhood filter can be seen as a partial differential equation (Buades, Coll, Morel [4]). With this interpretation, the depth difference constraints in the discrete case can be understood as the Dirichlet boundary conditions. Furthermore they allow to handle multiple depth layers.

Figure 11 is an example of the diffusion through the DDF. Using Eq. (4) a very large number of iterations are needed to attain the stability. To make the diffusion faster, the following equation is used as initialization:

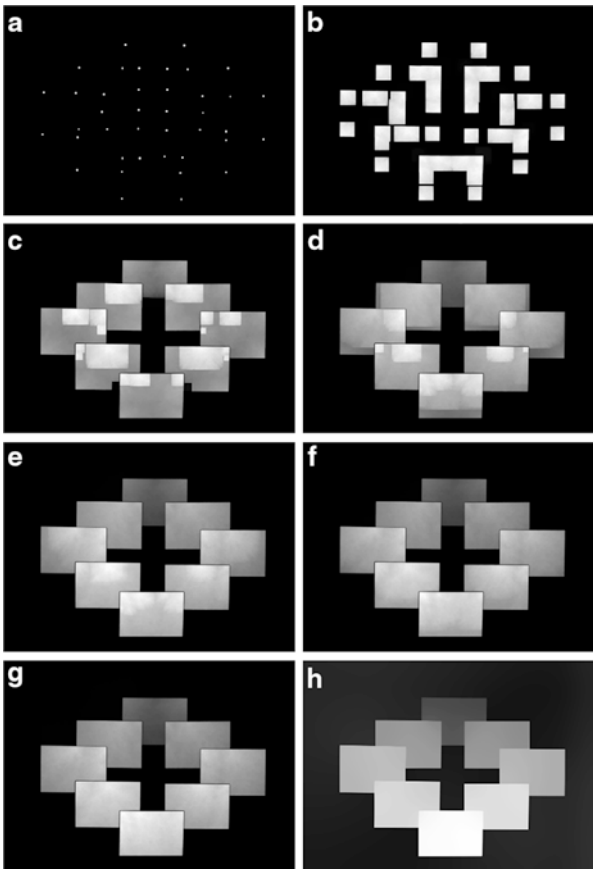


Fig. 12 Example of depth diffusion using Eq. (6) to speed up the diffusion. (**a**) Depth image, where FSPs have been initialized with a positive value (marked in *gray*) and the rest of the image with value zero. From (**b**) to (**g**) depth images corresponding to an increasing number of iterations. After each iteration, the depth difference between corresponding FSPs and BSPs is forced to be at least equal to the initial depth difference Δ, by adding Δ to FSPs when the difference between corresponding FSPs and BSPs becomes less than Δ. (**h**) Final depth image obtained using Eq. (4) on image (**g**)

$$\mathrm{DDF}_{h,r} z(x) = \sup_{y \in S_r(x)} z(y) e^{\frac{-|u(x)-u(y)|^2}{h^2}}, \tag{6}$$

while Eq. (4) is used only in the last iterations (see Fig. 12).

Experimental results on real images (see [10]) proved that this simple formulation turns out to be very effective for the integration of several monocular depth cues. In particular, contradictory information given by conflicting depth cues is dealt with the bilateral diffusion mechanism, which allows two regions to invert harmoniously their depths, in full agreement with the phenomenology.

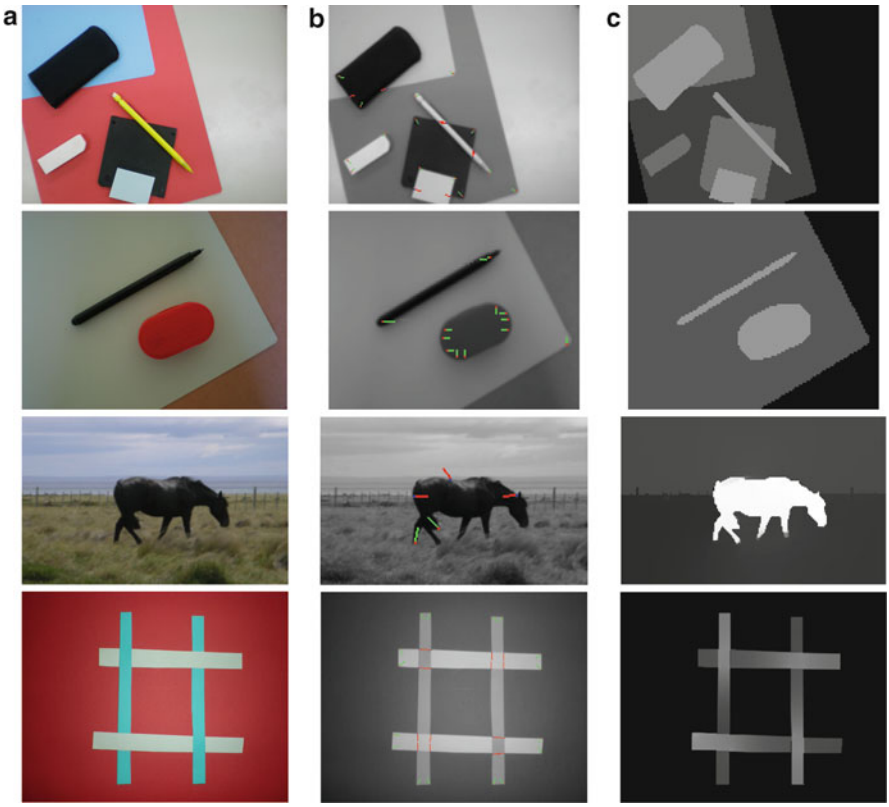


Fig. 13 (**a**) Original image. (**b**) Local depth cues are represented through vectors that point to the region closer to the viewpoint. (**c**) Depth image

In Fig. 13, some experimental result involving occlusion and convexity is shown. For each experiment three images are shown.

First, the original image (Fig. 13a) is shown. Then, on the second image, the initial depth gradient at depth cue points is represented by vectors pointing to the region closer to the viewpoint (red vectors arise from T-junctions, green vectors arise from local convexity) (Fig. 13b). Finally, the third image is the final result of the bilateral diffusion method (Fig. 13c). In this depth map high values indicate regions that are close to the camera. First and second rows of Fig. 13 show examples of indoor scenes, for which a proper solution is obtained. On the third row, there is an example of an outdoor scene involving a conflict. The T-junction detected on the back of the horse is due to a reflectance discontinuity, and its local depth interpretation is incorrect. However, on the depth map, the shape of the horse appears clearly on the foreground since the diffusion process allowed to overcome the local inconsistency. On the last row there is an example involving self-occlusion: occluding contours have different depth relationships at different points along its

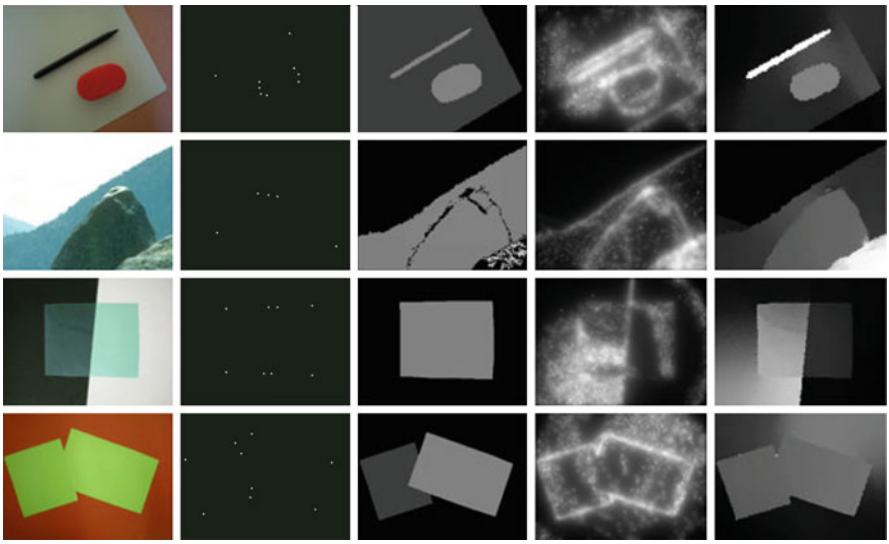


Fig. 14 (*1st column*) Original image. (*2nd column*) Depth cues computed by Dimiccoli, Morel, and Salembier [11]: FSPs are initialized with a positive value (marked in *white*), and BSP and the rest of the images are initialized with value zero. (*3rd column*) Depth image computed by Dimiccoli, Morel, and Salembier [11]. (*4th column*) Local depth cues computed by Calderero and Caselles [5]: they are encoded through gray-level values. High values indicate pixels closer to the viewpoint. (*5th column*) Depth image computed by Calderero and Caselles [5]

continuum. However, the bilateral diffusion method performs also well in this ambiguous situation.

More recently, the idea of using a neighborhood filter to globally integrate local depth information in single images has been re-proposed by Calderero and Caselles [5]. With the aim of achieving more accurate diffusion results on real images, they relied on color image information to determine the adaptive neighborhood by using the bilateral filter of Tomasi and Manduchi [37]. In other words, they averaged depth information on a pixel neighborhood using color-based weights in the same spirit as Yang et al. [42] and Kopf et al. [24].

Contrary to Dimiccoli, Morel, and Salembier [11], the local depth information extracted by Calderero and Caselles [5] is not sparse but behaves in a continuous manner. More precisely, the depth value at each pixel encodes the likelihood that an occlusion process is taking place at that pixel (see the fourth column of Fig. 14). This information is computed in a multi-scale fashion and therefore has the advantage of being more robust with respect to the depth information extracted by Dimiccoli, Morel, and Salembier [11]. However, due to its local nature, their method has the inconvenient of not being able to handle special cases of occlusions that involve more global processes, such as transparency and amodal completion, as illustrated in Fig. 14.

Due to the fact that depth information is initially estimated for each image pixel, Calderero and Caselles [5] use the neighborhood filter solely for assuring depth homogeneity of neighbor regions having similar color. Their diffusion filter is simpler and more robust with respect to errors coming from the estimation of local depth cues, but contrary to the filter of Dimiccoli, Morel, and Salembier [11], it does not allow to handle the case of multiple depth layers.

In conclusion, the method of Calderero and Caselles [5] gives better results on real images, whereas the method of Dimiccoli, Morel, and Salembier [11] gives results that better accord to phenomenology.

4 Conclusion

This chapter has reviewed the use of neighborhood filters for the recovery of 3D information. The first part of the chapter has reviewed bilateral filters applied to 3D data point sets, often organized in a triangulation, and has compared them to the simplest possible isotropic filter, the mean curvature motion, illustrating the advantage of bilateral filters on isotropic filtering. The second part of the chapter has reviewed bilateral filter applied to a data composed of a coarse, often sparse, depth map (or of depth cues) obtained through stereoscopy, or monocular depth cues estimation techniques. When diffusing depth information with a bilateral filter based on locality and color similarity, the discontinuities in depth are assured to be consistent with the color discontinuities, which is generally a desirable property. Experimental results have shown that the bilateral filter used as a diffusion tool performs well in restoring a dense depth map.

Acknowledgments The David raw point set is courtesy of the Digital Michelangelo Project, Stanford University. Fragment "31u" is courtesy of the Stanford Forma Urbis Romae Project, Stanford University and the Sovraintendenza of Rome. The Screw Nut point set is provided by the AIM@SHAPE repository and is courtesy of Laurent Saboret, INRIA. Research is partially financed by Institut Farman, ENS Cachan, the Centre National d'Etudes Spatiales (MISS Project), the European Research Council (advanced grant Twelve Labours), and the Office of Naval research (grant N00014-97-1-0839).

Cross-References

▸ Local Smoothing Neighborhood Filters
▸ Manifold Intrinsic Similarity
▸ Total Variation in Imaging

References

1. Adams, A., Gelfand, N., Dolson, J., Levoy, M.: Gaussian kd-trees for fast high-dimensional filtering. ACM Trans. Graph. **28**(3), 1–12 (2009)
2. Ansar, A., Castano, A., Matthies, L.: Enhanced real-time stereo using bilateral filtering. In: Proceedings of the 3D Data Processing, Visualization, and Transmission, 2nd International Symposium (3DPVT '04), Washington, DC, pp. 455–462. IEEE Computer Society (2004)

3. Buades, A., Coll, B., Morel, J.M.: A review of image denoising algorithms, with a new one. Multiscale Model. Simul. **4**(2), 490–530 (2005)
4. Buades, T., Coll, B., Morel, J.-M.: Neighborhood filters and pde's. Numerische Mathematik **105**(1), 11–34 (2006)
5. Calderero, F., Caselles, V.: Recovering relative depth from low-level features without explicit t-junction detection and interpretation. Int. J. Comput. Vis. **104**(1), 38–68 (2013)
6. Choudhury, P., Tumblin, J.: The trilateral filter for high contrast images and meshes. In: ACM SIGGRAPH 2005 Courses (SIGGRAPH '05), Los Angeles, p. 5. ACM, New York (2005)
7. Delage, E., Lee, H., Ng, Y.: A dynamic bayesian network model for autonomous 3d reconstruction from a single indoor image. In: International Conference on Computer Vision and Pattern Recognition (CVPR), New York, pp. 1–8 (2006)
8. Delon, J., Rougé, B.: Small baseline stereovision. J. Math. Imaging Vis. **28**(3), 209–223 (2007)
9. Digne, J., Morel, J.-M., Mehdi-Souzani, C., Lartigue, C.: Scale space meshing of raw data point sets. Comput. Graph. Forum, **30**(6), 1630–1642 (2011)
10. Dimiccoli, M.: Monocular depth estimation for image segmentation and filtering. PhD thesis, Technical University of Catalonia (UPC) (2009)
11. Dimiccoli, M., Morel, J.M., Salembier, P.: Monocular depth by nonlinear diffusion. In: Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP), Bhubaneswar, Dec 2008
12. Dunn, J.C.: A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. J. Cybern. **3**(3), 32–57 (1973)
13. Facciolo, G., Caselles, V.: Geodesic neighborhoods for piecewise affine interpolation of sparse data. In: International Conference on Image Processing, Cairo (2009)
14. Facciolo, G., Lecumberry, F., Almansa, A., Pardo, A., Caselles, V., Rougé, B.: Constrained anisotropic diffusion and some applications. In: British Machine Vision Conference, Edinburgh (2006)
15. Fleishman, S., Drori, I., Cohen-Or, D.: Bilateral mesh denoising. ACM Trans. Graph. **22**(3), 950–953 (2003)
16. Gamble, E., Poggio, T.: Visual integration and detection of discontinuities: the key role of intensitiy edges. Technical report 970, MIT AI Lab Memo (1987)
17. Gehrig, S.K., Franke, U.: Improving stereo sub-pixel accuracy for long range stereo. In: Proceedings of the 11th International Journal of Computer Vision, pp. 1–7 (2007)
18. Hoiem, D., Stein, A.N., Efros, A.A., Hebert, M.: Recovering occlusion boundaries from a single image. In: Proceedings of the International Conference on Computer Vision (ICCV), Rio de Janeiro, Brazil, pp. 1–8 (2007)
19. Hou, Q., Bai, L., Wang, Y.: Mesh smoothing via adaptive bilateral filtering. In: Springer (ed.) Computational Science – ICCS 2005, Atlanta, pp. 273–280. Springer (2005)
20. Jones, T.R., Durand, F., Desbrun, M.: Non-iterative, feature-preserving mesh smoothing. In: ACM SIGGRAPH 2003 Papers (SIGGRAPH '03), San Diego, pp. 943–949. ACM, New York (2003)
21. Jones, T.R., Durand, F., Zwicker, M.: Normal improvement for point rendering. IEEE Comput. Graph. Appl. **24**(4), 53–56 (2004)
22. Kellman, P.J., Shipley, T.F.: Visual interpolation in object perception. Curr. Dir. Psychol. Sci. **1**(6), 193–199 (1991)
23. Koller, D., Trimble, J., Najbjerg, T., Gelfand, N., Levoy, M.: Fragments of the city: Stanford's digital forma urbis romae project. In: Proceedings of the Third Williams Symposium on Classical Architecture, Rome. Journal of Roman Archaeology Supplementary, vol. 61, pp. 237–252 (2006)
24. Kopf, J., Cohen, M., Lischinski, D., Uyttendaele, M.: Joint bilateral upsampling. ACM Trans. Graph. **25**(3), 38–68 (2007)
25. Lafarge, F., Descombes, X., Zerubia, J., Pierrot-Deseilligny, M.: Automatic building extraction from dems using an object approach and application to the 3d-city modeling. J. Photogramm. Remote Sens. **63**(3), 365–381 (2008)

26. Lee, J.-S.: Digital image smoothing and the sigma filter. Comput. Vis. Graph. Image Process. **24**(2), 255–269 (1983)
27. Liu, Y.-S., Yu, P.-Q., Yong, J.-H., Zhang, H., Sun, J.-G.: Bilateral filter for meshes using new predictor. In: Springer (ed.) Computational and Information Science. Lecture Notes in Computer Science, vol. 3314/2005, pp. 1093–1099. Springer, Heidelberg (2005)
28. Mattoccia, S., Giardino, S., Gambin, A.: Accurate and efficient cost aggregation strategy for stereo correspondence based on approximated joint bilateral filtering. In: Asian Conference on Computer Vision (ACCV09), Xi'an (2009)
29. Metzger, W.: Gesetze des Sehens. Waldemar, Kramer (1975)
30. Miropolsky, A., Fischer, A.: Reconstruction with 3d geometric bilateral filter. In: Proceedings of the Ninth ACM Symposium on Solid Modeling and Applications (SM '04), Genoa, pp. 225–229. Eurographics Association, Aire-la-Ville (2004)
31. Ohtake, Y., Belyaev, A.G., Seidel, H.-P.: Mesh smoothing by adaptive and anisotropic Gaussian filter applied to mesh normals. In: VMV, Erlangen, pp. 203–210 (2002)
32. Paris, S., Durand, F.: A fast approximation of the bilateral filter using a signal processing approach. Int. J. Comput. Vis. **81**(1), 24–52 (2009)
33. Paris, S., Kornprobst, P., Tumblin, J., Durand, F.: Bilateral Filtering: Theory and Applications. Found. Trends Comput. Graph. Vis. **4**(1), 1–73 (2008). Hanover (2009)
34. Rother, D., Sapiro, G.: Seeing 3d objects in a single 2D image. In: 2009 IEEE 12th International Conference on Computer Vision (ICCV), Kyoto, pp. 1819–1826 (2009)
35. Sabater, N.: Reliability and accuracy in stereovision. Application to aerial and satellite high resolution image. PhD thesis, ENS Cachan (2009)
36. Scharstein, D., Szeliski, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. Int. J. Comput. Vis. **47**(1–3), 7–42 (2002)
37. Tomasi, C., Manduchi, R.: Bilateral filtering for gray and color images. In: Proceedings of the Sixth International Conference on Computer Vision (ICCV '98), Bombay, p. 839. IEEE Computer Society, Washington, DC (1998)
38. Wang, C.: Bilateral recovering of sharp edges on feature-insensitive sampled meshes. IEEE Trans. Vis. Comput. Graph. **12**(4), 629–639 (2006)
39. Wang, L., Liao, M., Gong, M., Yang, R., Nistér, D.: High-quality real-time stereo using adaptive cost aggregation and dynamic programming. In: Third International Symposium on 3D Data Processing, Visualization and Transmission (3DPVT), Chapel Hill (2006)
40. Wang, L., Yuan, B., Chen, J.: Robust fuzzy c-means and bilateral point clouds denoising. In: 2006 8th International Conference on Signal Processing, Beijing, vol. 2, pp. 16–20 (2006)
41. Wang, R.-F., Zhang, S.-Y., Zhang, Y., Ye, X.-Z.: Similarity-based denoising of point-sampled surfaces. J. Zhejiang Univ. **9**(6), 807–815 (2008)
42. Yang, Q., Wang, L., Yang, R., Stewénius, H., Nistér, D.: Stereo matching with color-weighted correlation, hierarchical belief propagation and occlusion handling. IEEE Trans. Pattern Anal. Mach. Intell. **31**(3), 1–13 (2006)
43. Yang, Q., Yang, R., Davis, J., Nistér, D.: Spatial-depth super resolution for range images. In: International Conference on Computer Vision and Pattern Recognition (CVPR), Minneapolis (2007)
44. Yaroslavsky, L.P.: Digital Picture Processing. An Introduction. Springer Series in Information Sciences, vol. 9 . Springer, Berlin/Heidelberg (1985)
45. Yin, J., Cooperstock, J.R.: Improving depth maps by nonlinear diffusion. In: Proceedings of 12th International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision (WSCG), Plzen-Bory, pp. 1–8 (2004)
46. Yoon, K.-J., Kweon, S.: Adaptive support-weight approach for correspondence search. IEEE Trans. Pattern Anal. Mach. Intell. **28**(4), 650–656 (2006)
47. Yoshizawa, S., Belyaev, A., Seidel, H.-P.: Smoothing by example: mesh denoising by averaging with similarity-based weights. In: Proceedings of the IEEE International Conference on Shape Modeling and Applications 2006 (SMI'06), Matsushima, p. 9. IEEE Computer Society, Washington, DC (2006)

Splines and Multiresolution Analysis

Brigitte Forster

Contents

1 Introduction 1676
2 Historical Notes 1681
3 Fourier Transform, Multiresolution, Splines, and Wavelets 1682
Mathematical Foundations 1682
B-Splines 1690
Polyharmonic B-Splines 1694
4 Survey on Spline Families 1696
Schoenberg's B-Splines for Image Analysis: The Tensor Product Approach 1696
Fractional and Complex B-Splines 1697
Polyharmonic B-Splines and Variants 1701
Splines on Other Lattices 1704
5 Numerical Implementation 1708
6 Open Questions 1711
7 Conclusion 1713
Cross-References 1713
References 1714

Abstract

Splines and multiresolution are two independent concepts, which – considered together – yield a vast variety of bases for image processing and image analysis. The idea of a multiresolution analysis is to construct a ladder of nested spaces that operate as some sort of mathematical looking glass. It allows to separate coarse parts in a signal or in an image from the details of various sizes. Spline functions are piecewise or domainwise polynomials in one dimension (1D) resp. nD. There is a variety of spline functions that generate multiresolution analyses. The viewpoint in this chapter is the modeling of such spline functions in frequency

B. Forster (✉)
Fakultät für Informatik und Mathematik, Universität Passau, Passau, Germany
e-mail: brigitte.forster@uni-passau.de

© Springer Science+Business Media New York 2015

O. Scherzer (ed.), *Handbook of Mathematical Methods in Imaging*,
DOI 10.1007/978-1-4939-0790-8_28

domain via Fourier decay to generate functions with specified smoothness in time domain resp. space domain. The mathematical foundations are presented and illustrated at the example of cardinal B-splines as generators of multiresolution analyses. Other spline models such as complex B-splines, polyharmonic splines, hexagonal splines, and others are considered. For all these spline families exist fast and stable multiresolution algorithms which can be elegantly implemented in frequency domain. The chapter closes with a look on open problems in the field.

AMS Subject Classification (2010): 41A15 Spline approximation 65D07 Numerical analysis – Splines 68U10 computing methodologies and applications – Image processing 65T99 numerical methods in Fourier analysis

1 Introduction

This chapter deals with two originally independent concepts, which were recently combined and since together have a strong impact on signal and image analysis: the concept of splines, i.e., of piecewise polynomials, and the concept of multiresolution analysis, i.e., splitting functions – or more general data – into coarse approximations and details of various sizes. Ever since the combination of the two concepts, they have led to a load of new applications in, e.g., signal and image analysis, as well as in signal and image reconstruction, computer vision, numerics of partial differential equations, and other numerical fields. An impression of the vast area of applications can be gained in, e.g., [1, 17, 20, 64].

Already, the spline functions alone proved to be very useful for mathematical analysis as well as for signal and image processing, analysis and representation, computer graphics, and many more; see, e.g., [3, 15, 22, 24, 35, 50, 58]. An example for a family of spline functions are I. J. Schoenberg's polynomial splines with uniform knots [59, 60]:

$$\beta^m(t) = \frac{1}{m!} \sum\nolimits_{k=0}^{m+1} (-1)^k \binom{m+1}{k} (t-k)_+^m, \quad m \in \mathbb{N}. \tag{1}$$

Here, t_+^m denotes the one-sided power function, i.e., $t_+^m = 0$ for $t < 0$ and $t_+^m = t^m$ for $t \geq 0$. The B-splines β^m can be easily generated by an iterative process. Let $\beta^0(t) = \chi_{[0,1)}(t)$ be the characteristic function of the interval $[0, 1)$. Then the B-spline of degree m is derived by the convolution product

$$\beta^m = \beta^0 * \beta^{m-1} = \underbrace{\beta^0 * \ldots * \beta^0}_{m+1\text{-times}} \quad \text{for } m \in \mathbb{N}, \tag{2}$$

where

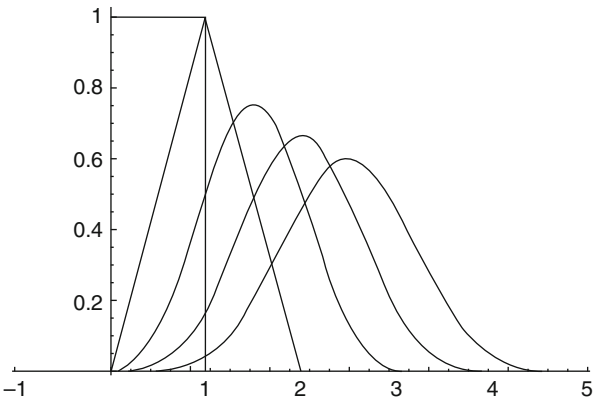


Fig. 1 Cardinal B-splines of degree $m = 0, \ldots, 4$

$$\beta^0 * \beta^{m-1}(x) = \int_{\mathbb{R}} \beta^0(y)\beta^{m-1}(x-y)dy = \int_0^1 \beta^{m-1}(x-y)dy.$$

For an illustration of the cardinal B-splines, see Fig. 1.

Splines had their second breakthrough as Battle [4] and Lemarié [38] discovered that B-splines generate multiresolution analyses. The simple form of the B-splines and their compact support, in particular, were convenient for designing multiresolution algorithms and fast implementations.

Definition 1. Let $A : \mathbb{R}^n \to \mathbb{R}^n$ be a linear mapping that leaves $\mathbb{Z}^n$ invariant, i.e., $A(\mathbb{Z}^n) \subset \mathbb{Z}^n$ and that has (real or complex) eigenvalues with absolute values greater than 1.

A multiresolution analysis associated with the dilation matrix A is a sequence of nested subspaces $(V_j)_{j\in\mathbb{Z}}$ of $L^2(\mathbb{R}^n)$ such that the following conditions hold:

(i) $\ldots \subset V_{-1} \subset V_0 \subset V_1 \subset \ldots$,
(ii) $\cap_{j\in\mathbb{Z}} V_j = \{0\}$,
(iii) Span $\cup_{j\in\mathbb{Z}} V_j$ is dense in $\mathrm{L}^2(\mathbb{R}^n)$,
(iv) $f \in V_j \Leftrightarrow f(A^{-j}\,\cdot) \in V_0$,
(v) $f \in V_0 \Leftrightarrow f(\cdot - k) \in V_0$ for all $k \in \mathbb{Z}^n$.
(vi) There exists a so-called scaling function $\phi \in V_0$ such that the family $\{\phi(\bullet - k)\}_{k\in\mathbb{Z}^n}$ of translates of ϕ forms a Riesz basis of V_0.

Here, $L^2(\mathbb{R}^n)$ denotes the vector space of square-integrable functions $f\colon \mathbb{R}^n \to \mathbb{C}$ with norm

$$\|f\|_2 = \left(\int_{\mathbb{R}^n} |f(x)|^2 dx\right)^{\frac{1}{2}}$$

and corresponding inner product

$$\langle f, g \rangle = \int_{\mathbb{R}^n} f(x)\overline{g(x)}\, dx,$$

where $\bar{g}$ denotes the complex conjugate of g. The elements in $L^2(\mathbb{R}^n)$ are also called functions of finite energy.

Riesz bases are a slightly more general concept than orthonormal bases. In fact, Riesz bases are equivalent to orthonormal bases and therefore can be generated by applying a topological isomorphism on some orthonormal basis.

Definition 2. A sequence of functions $\{f_n\}_{n\in\mathbb{Z}}$ in a Hilbert space V is called a Riesz sequence if there exist positive constants A and B, the Riesz bounds, such that

$$A\|c\|_{l^2} \leq \left\| \sum\nolimits_{k\in\mathbb{Z}^n} c_k f_k \right\|_V \leq B\|c\|_{l^2}$$

for all scalar sequences $C = (c_k)_{k\in\mathbb{Z}^n} \subset l^2(\mathbb{Z}^n)$.

A Riesz sequence is called a Riesz basis if it additionally spans the space V.

A good introduction to Riesz bases, their properties, and their relation to orthonormal bases is given in the monography by Young [75]. Multiresolution constructions with splines are treated in numerous sources. As a starting point, there are, e.g., the books by Christensen [13, 14] and Wojtaszczyk [74].

The mathematical properties in Definition 1 have intuitive interpretations. A function $f \in L^2(\mathbb{R}^n)$, which is projected orthogonally on V_j, is approximated with the so-called resolution A^j. In fact, let

$$P_j : L^2(\mathbb{R}^n) \to V_j$$

denote the orthogonal projection operator. Then (ii) yields that by going to lower resolutions, all details are lost:

$$\lim_{j\to-\infty} \left\| P_j f \right\| = 0.$$

In contrast, when the resolution is increased, $j \to \infty$, more and more details are added. By (iii), the projection then converges to the original function f:

$$\lim_{j\to\infty} \left\| f - P_j f \right\| = 0.$$

Hereby, the rate of convergence depends on the regularity of f.

The approximation spaces V_j are nested, which allows for computing coarser approximations in V_k for $k < j$ for functions $f \in V_j$. The scaling A^k enlarges details. Property (iv) shows that the approximation spaces have a similar structure over the scales and emanate from one another. The translation invariance (v) ensures that the analysis of a function in V_j is independent of the starting time or location.

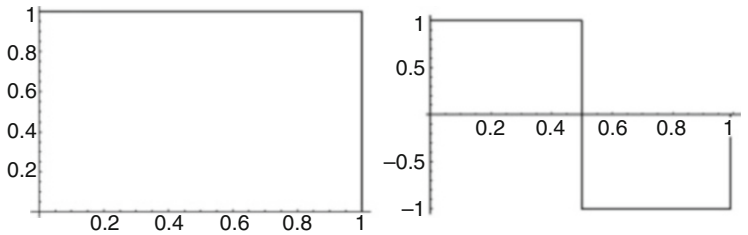


Fig. 2 *Left*: The B-spline β^0 is a scaling function and operates as mean value. *Right*: The corresponding wavelet, the Haar wavelet, operates as a local difference operator

And property (vi) finally ensures the beautiful and mighty property that the whole sequence of nested approximation spaces can be generated by translates and scalings of one single function – the scaling function. In fact, (vi) together with (iv) yields that

$$\{\phi(A^j \bullet -k), k \in \mathbb{Z}^n\}$$

is a Riesz basis for V_j.

While moving from the coarser approximation space V_j to the finer, larger space V_{j+1}, information has to be added. In fact, there is an orthogonal complement W_j, $j \in \mathbb{Z}$, such that

$$V_{j+1} = V_j \oplus W_j.$$

These spaces are called detail spaces or wavelet spaces. It is well known that these spaces also possess a Riesz basis spanned by shifts of $|\det A| - 1$ generators, the wavelets $\psi_1, \ldots, \psi_{|\det A|-1}$. Here, A is the dilation matrix in Definition 1. The wavelets can be constructed from the scaling function. As a consequence, the knowledge of just the single function ϕ allows for the construction of the approximation spaces V_j and for the wavelet spaces W_j. Detailed information on the generation of wavelets and their properties can be found in various books, e.g., [16, 21, 41, 44, 74].

Example 1. A simple example for a multiresolution analysis on $L^2(\mathbb{R})$ is given by piecewise constant functions. Consider the characteristic function $\phi = \chi_{[0,1)}$ of the interval semi-open interval [0, 1). Then ϕ generates a dyadic multiresolution analysis, i.e., for $A = 2$. The approximation spaces are

$$V_j = \overline{\text{span}\,\{\chi_{[0,1)}(2^j \bullet -k)\}_{k\in\mathbb{Z}}}^{L^2(\mathbb{R})}.$$

They consist of functions constant on intervals of the form $[k2^{-j}, (k+1)2^{-j})$. The spaces are obviously nested and separate $L^2(\mathbb{R})$ in the sense of Definition 1 (ii).

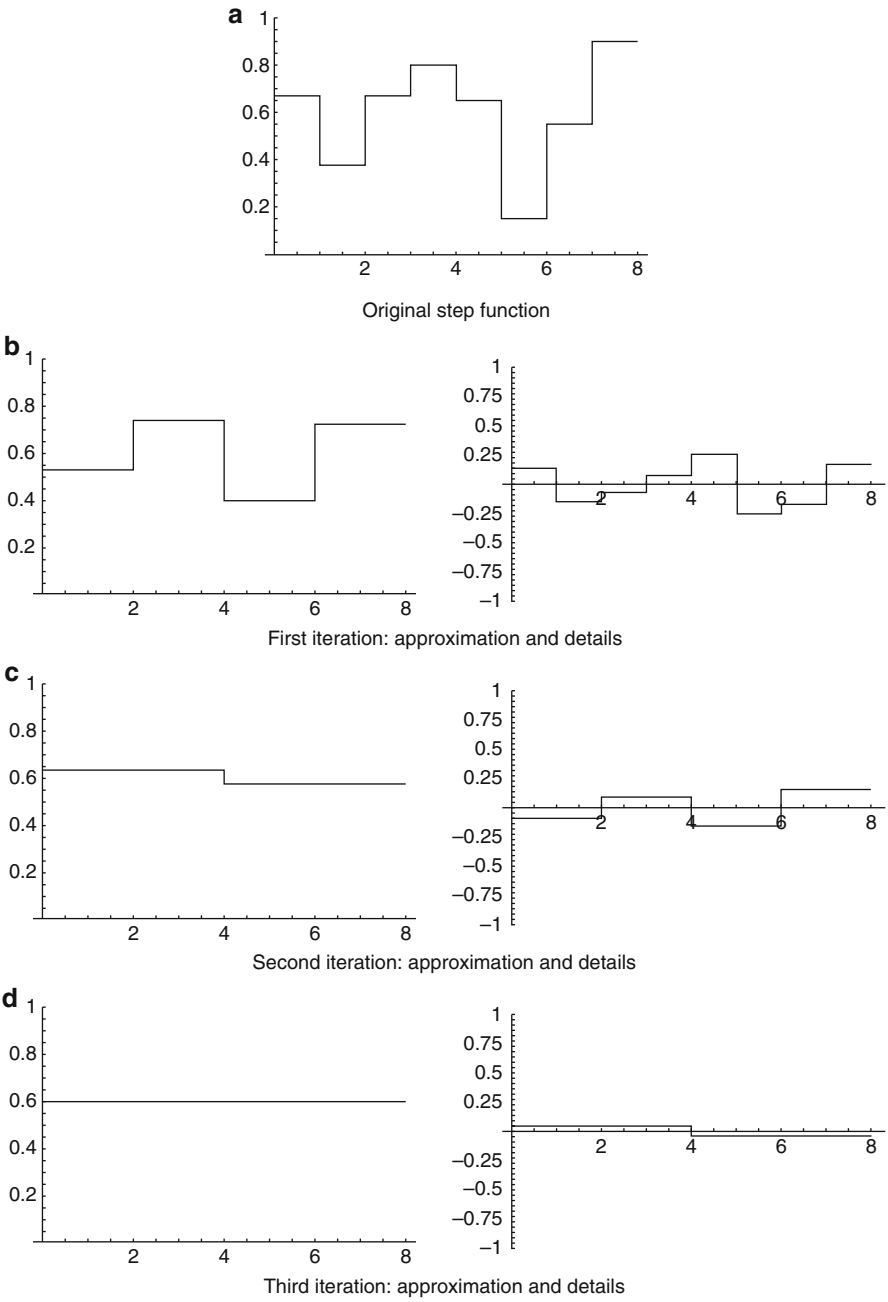


Fig. 3 Multiresolution decomposition of the step function (**a**) with β^0 as scaling function and with the Haar wavelet. The approximations (*left column*) are further iterated and decomposed into a coarser approximation and details (*right column*), until the coarsest approximation step, here the mean value, is reached. The sum of the coarsest approximation in the third iteration and of all details yields the original function (**a**). No information is lost in a multiresolution decomposition

Since piecewise constant functions with compact support are dense in $L^2(\mathbb{R})$, (iii) holds. (iv) – (vi) hold by construction. In fact, this multiresolution analysis is generated by the B-spline β^0 as scaling function. The B-spline basis operates as mean-value operator over the support interval. The corresponding wavelet extracts the details, i.e., the deviation from the mean value. To this end, it operates as a difference operator. Figure 2 shows the scaling function β^0 and the corresponding wavelet, the so-called Haar wavelet. In Fig. 3, an example of a multiresolution is given.

2 Historical Notes

The idea of piecewise polynomial functions and splines goes back to Schoenberg [59, 60]. In the 1960s, when computer power started to be used for numerical procedures such as data fitting, interpolation, solving differential equations, and computer-aided geometric design, splines experienced an extreme upsurge. Schoenberg invented and strongly contributed to the concept of cardinal splines, which have equidistant nodes on the integers; see, e.g., [40, 61, 62] and many more.

As a parallel development, in the 1980s, the adaption of signal resolution to only process relevant details for a particular task evolved. For example, for computer vision, a multiresolution pyramid was introduced by Burt and Adelson [8]. It allowed to process an image first on a low-resolution level and then selectively increase the resolution and add more detailed information when needed. The definition of a dyadic multiresolution analysis, i.e., $A = 2Id$, was contributed by Mallat [43] and Meyer [46]. An interesting and in some parts historical collection on the most important articles in multiresolution and wavelet theory was assembled by Heil and Walnut [33].

The concepts of splines and multiresolution were joined by Lemarié [38] and Battle [4], when they showed that cardinal B-splines are scaling functions for multiresolution analyses. This led to many more developments of piecewise polynomial scaling functions for various settings and also multidimensions [15], as, e.g., polyharmonic B-splines [53, 54] and other functions inspired from radial basis functions [7].

In 1989, S. Mallat published his famous algorithm for multiresolution and wavelet analysis [43]. He had developed an efficient numerical method such that multiresolution decompositions could be calculated in a fast way. For the splines, M. Unser et al. proposed a fast implementation [66,68,69] which strongly contributed to the breakthrough of splines for signal and image analysis. In the last years, periodic, fractional, and complex versions of splines for multiresolution were developed, e.g., [11, 27, 28, 51, 65]. Many of them use a Fourier domain filter algorithm which allows for infinite impulse response filters. The former important feature of compact support of the cardinal B-splines and other functions is no longer a limiting criterion. Therefore, it can be expected that many new contributions on splines will still be made in the future by modeling signal and image features in Fourier domain.

3 Fourier Transform, Multiresolution, Splines, and Wavelets

Mathematical Foundations

Regularity and Decay Under the Fourier Transform

An important idea behind splines and multiresolution is the relation between regularity in time domain and decay in frequency domain and, respectively, between decay in time domain and regularity in frequency domain. To illustrate this, the notion of the Schwartz space is very useful [10, 34, 56, 63].

Definition 3. The subspace of functions $f \in C^\infty(\mathbb{R}^n)$ with

$$\sup_{|\alpha|\le N} \sup_{x\in\mathbb{R}^n} (1+\|x\|^2)^N |D^\alpha f(x)| \;<\; \infty \quad \text{for all } N = 0,\ 1,\ 2,\ \ldots$$

is called the space of rapidly decreasing functions or Schwartz space $\mathcal{S}(\mathbb{R}^n)$. The norms induce a Fréchet space topology, i.e., the space $\mathcal{S}(\mathbb{R}^n)$ is complete and metrizable.

Here, $D^\alpha = \left(\frac{\partial}{\partial x_1}\right)^{\alpha_1} \cdots \left(\frac{\partial}{\partial x_n}\right)^{\alpha_n}$ for every multi-index $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{N}_o^n$.

The dual space $\mathcal{S}'(\mathbb{R}^n)$, endowed with the weak-* topology, is called space of tempered distributions.

The following famous linear transform relates the viewpoints of the space domain and of the frequency domain:

Definition 4. The Fourier transform, defined by

$$\mathcal{F}f(\omega) := \hat{f}(\omega) := \int_{\mathbb{R}^n} f(x) e^{-i\langle\omega,x\rangle} dx, \quad \omega \in \mathbb{R}^n,$$

is a topological isomorphism on $L^2(\mathbb{R}^n)$ and on $\mathcal{S}(\mathbb{R}^n)$. Its inverse is given by

$$f(x) = \frac{1}{(2\pi)^n} \int_{\mathbb{R}^n} \hat{f}(\omega) e^{i\langle\omega,x\rangle} d\omega \quad \text{in } L^2(\mathbb{R}^n) \text{ resp. in } \mathcal{S}(\mathbb{R}^n).$$

The Fourier transform can be extended to the space of tempered distributions. For $T \in \mathcal{S}'(\mathbb{R}^n)$, the Fourier transform is defined in a weak sense as

$$\mathcal{F}T(\phi) := \hat{T}(\phi) := T(\hat{\phi}) \text{ for all } \phi \in \mathcal{S}(\mathbb{R}^n).$$

Also on $\mathbb{S}'(\mathbb{R}^n)$, the Fourier transform is a topological isomorphism.

The Fourier transform has the nice property to relate polynomials and differential operators.

Theorem 1.

(i) Let $f \in \mathcal{S}(\mathbb{R})$. Then for all $k \in \mathbb{N}$

$$\mathcal{F}\left(f^{(k)}\right)(\omega) = (i\omega)^k \hat{f}(\omega),$$

and

$$\hat{f}^{(k)}(\omega) = \mathcal{F}\left((-i\bullet)^k f\right)(\omega).$$

(ii) Let P be an algebraic polynomial in $\mathbb{R}^n$, say $P(x) = \sum_\alpha c_\alpha x_1^{\alpha_1} \dots x_n^{\alpha_n}$, and let $f \in \mathcal{S}(\mathbb{R}^n)$. Then

$$\mathcal{F}\left(P\left(\frac{1}{i}D\right) f\right) = P\,\hat{f} \quad \text{and} \quad \widehat{Pf} = P(iD)\,\hat{f},$$

where $P(iD) = \sum_\alpha c_\alpha i^{|\alpha|} D^\alpha$.

(iii) Part (ii) also holds for $f \in \mathcal{S}'(\mathbb{R}^n)$.

Example 2. The Fourier transform of the polynomial x^k is the tempered distribution $i^k \frac{d^k}{dx^k}\delta$, $k \in \mathbb{N}_0$.

For the construction of a multiresolution analysis, the scaling function can be used as a starting point. The idea is to choose a scaling function of a certain regularity, such that the generated multiresolution analysis inherits the smoothness properties. In particular, for the splines, the idea is to model the regularity via decay in Fourier domain. The following theorem gives a motivation for this. The result can be deduced from the considerations above, and the fact that $\mathcal{S}(\mathbb{R}^n)$ is dense in $L^2(\mathbb{R}^n)$:

Theorem 2. *Let $f \in L^2(\mathbb{R}^n)$ and its Fourier-transform decay as*

$$|\hat{f}(\omega)| \leq C(1 + \| \omega \|)^{-N-\epsilon}$$

for some $\varepsilon > 0$. Then all partial derivatives of order $\leq N - n$ are continuous and in $L^2(\mathbb{R}^n)$.

These results allow to construct a scaling function with explicit regularity and decay properties, in space and in frequency domain. However, some criteria are needed to verify that the constructed function generates a multiresolution analysis.

Criteria for Riesz Sequences and Multiresolution Analyses

The following is an explicit criterion to verify whether some function ϕ is a scaling function.

Theorem 3. *Let A be a dilation matrix and let $\phi \in L^2(\mathbb{R}^n)$ be some function satisfying the following properties:*

(i) $\{\phi(\bullet - k)\}_{k\in\mathbb{Z}^n}$ *is a Riesz sequence in* $L^2(\mathbb{R}^n)$.
(ii) ϕ *satisfies a scaling relation. That is, there is a sequence of coefficients* $(a_k)_{k\in\mathbb{Z}^n}$ *such that*

$$\phi(A^{-1}x) = \sum\nolimits_{k\in\mathbb{Z}^n} a_k \phi(x+k) \quad \text{in } L^2(\mathbb{R}^n). \tag{3}$$

(iii) $|\hat{\phi}|$ *is continuous at 0 and* $\hat{\phi}(0) \neq 0$.

Then the spaces

$$V_j = \text{span}\,\{\phi(A^j \bullet - k)\}_{k\in\mathbb{Z}^n}, \quad j \in \mathbb{Z},$$

form a multiresolution analysis of $L^2(\mathbb{R}^n)$ *with respect to the dilation matrix* A.

Proof. See, e.g., [74, Theorem 2.13] for the 1D case.

For particular applications, the Riesz basis property (i) of $\{\phi(\bullet - k)\}_{k\in\mathbb{Z}^n}$ in V_0 is not enough, but an orthonormal basis is needed. An example for such an application is the denoising of signals contaminated with Gaussian white noise [44, Chap. X, Sect. 10.2.1]. However, there is an elegant mathematical method to orthonormalize Riesz bases generated by shifts of a single function.

Theorem 4. *Let* $\phi \in L^2(\mathbb{R}^n)$. *Then the following holds:*

(i) $\{\phi(\bullet - k)\}_{k\in\mathbb{Z}^n}$ *is a Riesz sequence in* $L^2(\mathbb{R}^n)$ *if and only if there are constants* c *and* C, *such that*

$$0 < c \leq \sum\nolimits_{k\in\mathbb{Z}^n} |\hat{\phi}(\omega + 2\pi k)|^2 \leq C < \infty \text{ almost everywhere.}$$

That is, the autocorrelation filter $M(\omega) := \sum_{k\in\mathbb{Z}^n} |\hat{\phi}(\omega + 2\pi k)|^2$ *is strictly positive and bounded from above.*
(ii) $\{\phi(\bullet - k)\}_{k\in\mathbb{Z}^n}$ *is an orthonormal sequence if and only if*

$$\sum\nolimits_{k\in\mathbb{Z}^n} |\hat{\phi}(\omega + 2\pi k)|^2 = 1 \text{ almost everywhere.}$$

(iii) If $\{\phi(\bullet - k)\}_{k\in\mathbb{Z}^n}$ *is a Riesz basis of a subspace* X *of* $L^2(\mathbb{R}^n)$, *then there exists a function* $\Phi \in L^2(\mathbb{R}^n)$, *namely,*

$$\hat{\Phi}(\omega) = \frac{\hat{\phi}(\omega)}{\sqrt{\sum_{k\in\mathbb{Z}^n} |\hat{\phi}(\omega + 2\pi k)|^2}} \tag{4}$$

such that $\{\Phi(\bullet - k)\}_{k\in\mathbb{Z}^n}$ *is an orthonormal basis of* X.

Proof. See, e.g., [74] and [44, Chap. VII].

Due to this theorem, every scaling function can be orthonormalized. Let $\phi \in L^2(\mathbb{R}^n)$ be some scaling function that generates a multiresolution analysis $\{V_j\}_{j\in\mathbb{Z}}$ of $L^2(\mathbb{R}^n)$. Then the family $\{\Phi_{j,k}\}_{k\in\mathbb{Z}^n}$ with

$$\Phi_{j,k}(x) = 2^{-j/2}\Phi(2^{-j}(x-k)),$$

and Φ as defined in (4) is an orthonormal basis of the space V_j, $j \in \mathbb{Z}$.

Example 3. A simple possibility to construct a dyadic multiresolution analysis in $L^2(\mathbb{R}^n)$ is the tensor product approach. Let $(V_j)_{j\in\mathbb{Z}}$ be a dyadic (i.e., $A = 2$) multiresolution analysis of $L^2(\mathbb{R})$ with scaling function ϕ. Then $(\boldsymbol{V}_j)_{j\in\mathbb{Z}}$ with $\boldsymbol{V}_j = \underbrace{V_j \otimes \cdots \otimes V_j}_{n-\text{times}}$ together with the scaling function $\boldsymbol{\phi}(x_1,\ldots,x_n) = \phi(x_1)\cdot\ldots\cdot\phi(x_n)$ forms a multiresolution analysis of $L^2(\mathbb{R}^n)$ and dilation matrix $2Id$.

In the same way, the scaling function $\boldsymbol{\phi}(x_1,\ldots,x_n) = \phi_1(x_1)\cdot\ldots\cdot\phi_n(x_n)$ generates a multiresolution analysis of $L^2(\mathbb{R}^n)$, if every $\phi_k, k = 1,\ldots,n$, is a scaling function of some 1D multiresolution analysis with dilation factor $a \in \mathbb{N}\setminus\{1\}$.

Regularity of Multiresolution Analysis

In signal and image analysis, the choice of an appropriate analysis basis is crucial. Here, appropriate means that the features of the basis such as smoothness should be in accordance with the properties of the functions to analyze. To give a blatant example: Analyzing a smooth signal or image with a fractal basis in general yields results that are difficult to interpret and to work with in practice. In this case, the signal resp. the image model does not match the model of the basis.

The next section will show that the family of spline bases helps to avoid such difficulties, because the splines allow for a good adjustment due to their regularity parameter m; cf. (1) and (2). The following definition specifies the term "regular." (See [74].)

Definition 5. Denote C^r the class of r-times continuously differentiable functions in $\mathbb{R}^n$, C^0 the class of continuous functions, and C^{-1} the class of measurable functions.

(i) Let $r = -1, 0, 1, \ldots$. A function $f : \mathbb{R}^n \to \mathbb{C}$ is called r-regular, iff $\in C^r$ and

$$\left|\frac{\partial^\alpha}{\partial x^\alpha} f(x)\right| \le \frac{A_k}{(1+\|x\|)^k}$$

for every $k \in \mathbb{N}_0$, every multi-index α with $|\alpha| \leq \max(r, 0)$ and constants A_k.

(ii) A multiresolution analysis of $L^2(\mathbb{R}^n)$ is called r-regular if it is generated by an r-regular scaling function.

It is important to note that the orthonormalization procedure (4) does not affect the regularity of the corresponding basis. For the orthonormalized scaling function Φ of a multiresolution analysis, the same regularity properties hold.

Proposition 1. *Let $\phi \in L^2(\mathbb{R}^n)$ be an r-regular function, such that $\{\phi(\bullet - k)\}_{k\in\mathbb{Z}^n}$ forms a Riesz sequence. Then the via (4) orthonormalized function Φ is also r-regular [74].*

Order of Approximation

Having found a scaling function that generates a multiresolution analysis, how good do the corresponding approximation spaces V_j approximate some function $f \in L^2(\mathbb{R}^n)$ of a certain regularity? Let

$$H^k(\mathbb{R}^n) = \{f \in L^2(\mathbb{R}^n) : \|f\|_{H^k} := \frac{1}{(2\pi)^n} \|(1+\|\bullet\|_{\mathbb{R}^n})^k \hat{f}\|_{L^2} < \infty\}, \quad k \in \mathbb{N}_0$$

denote the Sobolev spaces. The following criterion for the order of approximation turns out to be easy to verify for splines.

Theorem 5. *Let $\phi \in L^2(\mathbb{R}^n)$ satisfy the following properties [23, Theorem 1.15]:*

(i) $1/\hat{\phi}$ *is bounded on some neighborhood of the origin.*

(ii) Let B_ε be some open ball centered at the origin and let $E := B_\varepsilon + (2\pi\mathbb{Z}^n \setminus \{0\})$. For some $\alpha > k + n/2$, all derivatives of $\hat{\phi}$ of order $\leq \alpha$ are in $L^2(E)$.

(iii) $D^\gamma \hat{\phi}(\omega) = 0$ for all $|\gamma| < k$ and all $\omega \in 2\pi\mathbb{Z}^d \setminus \{0\}$.

Then $V_0 = \overline{\text{span}\,\{\phi(\bullet - k)\}_{k\in\mathbb{Z}^n}}$ provides an approximation order k: For $f \in H^k(\mathbb{R}^n)$,

$$\min\{\|f - s(\bullet/h)\|_{L^2}, s \in V_0\} \leq \text{const.}\, h^k \|f\|_{H^k} \quad \text{for all } h > 0.$$

Wavelets

For the step from a coarser approximation space V_j to a finer one V_{j+1}, information has to be added. It is contained in the wavelet space or detail space W_j, which is the orthonormal complement of V_j in V_{j+1}:

$$V_{j+1} = V_j \oplus W_j.$$

It follows that $V_{j+m} = V_j \oplus \bigoplus_{l=0}^{k-1} W_{j+l}$, and hence

$$L^2(\mathbb{R}^n) = \bigoplus_{j\in\mathbb{Z}} W_j \tag{5}$$

can be decomposed in a direct sum of mutually orthogonal detail spaces. Moreover, the detail spaces W_j inherit the scaling property from Definition 1(iv) for the approximation spaces V_j. For all $j \in \mathbb{Z}$,

$$f \in W_j \Leftrightarrow f(A^{-j}\bullet) \in W_0.$$

The question now is whether there is also a simple basis generated by the shifts of one or few functions, the wavelets. The following definition is motivated from Eq. (5).

Definition 6. Let A be a dilation matrix, and let $\{\psi_l\}_{l=1,\dots,q}, q \in \mathbb{N}$, be a set of functions in $L^2(\mathbb{R}^n)$, such that the family

$$\{|\det A|^{j/2}\psi_l(A^j \bullet -k)|\ l = 1,\dots,q,\ j \in \mathbb{Z},\ k \in \mathbb{Z}^n\}$$

forms an orthonormal basis of $L^2(\mathbb{R}^n)$. Then $\{\psi_l\}_{l=1,\dots,q}$ is called a wavelet set associated with A.

What qualitative properties do the wavelets have? The approximation spaces V_j are generated by the scaling function, which operates as a low-pass filter. This can be seen from Theorem 3(iii) $\hat{\phi}(0) \neq 0$ and resp. from Theorem 5(i): $1/\hat{\phi}$ is bounded in some neighborhood of the origin. Therefore, the added details and thus the wavelets have to carry the high-frequency information. In addition, the wavelets ψ in W_0 are elements of V_1 and therefore have the form

$$\psi(A^{-1}x) = \sum\nolimits_{k\in\mathbb{Z}^n} a_k \phi(x-k) \tag{6}$$

in L^2 norm, where $\{a_k\}_{k\in\mathbb{Z}^n}$ are the Fourier coefficients of a certain $2\pi\mathbb{Z}^n$-periodic function.

Proposition 2. *Let $(V_j)_{j\in\mathbb{Z}}$ be a multiresolution analysis of $L^2(\mathbb{R}^n)$ with respect to the dilation matrix A and with scaling function ϕ. Then for a function $f \in L^2(\mathbb{R}^n)$ the following are equivalent: $f \in V_1$ id and only if*

$$\hat{f}(A^T\omega) = m_f(\omega)\hat{\phi}(\omega) \text{ almost everywhere.}$$

Here, $m_f \in L^2([0,2\pi]^n)$ and

$$\|m_f\|^2_{L^2([0,2\pi]^n)} = \frac{1}{|\det A|}\|f\|^2_{L^2(\mathbb{R}^n)}.$$

For a proof, see, e.g., [21, 74]. Note that for a wavelet ψ as in Eq. (6), there holds

$$m_\psi(\omega) = \frac{1}{|\det A|} \sum\nolimits_{k\in\mathbb{Z}^n} a_k e^{i\langle\omega,k\rangle}.$$

How many wavelets, i.e., generators of W_0, are needed to span the space? The parameter q in Definition 6 is yet unspecified. In fact, q depends on the scaling matrix. A leaves the lattice $\mathbb{Z}^n$ invariant, $A\mathbb{Z}^n \subset \mathbb{Z}^n$. The number of cosets is $|\det A| = |\mathbb{Z}^n / A\mathbb{Z}^n|$ (see [74, Proposition 5.5]). It turns out that $q = |\det A| - 1$ wavelets are needed to generate the space W_0. To motivate this, for a start, let $f \in V_1$ be an arbitrary function. Denote $\gamma_0, \ldots, \gamma_q$ representatives of the $q+1$ cosets of $A\mathbb{Z}^n$ in $\mathbb{Z}^n$. Then each coset can be written as $\gamma_m + A\mathbb{Z}^n$, $m = 0, \ldots, q$. The function f has the representation

$$\frac{1}{|\det A|^{1/2}} f(A^{-1}x) = \sum\nolimits_{k\in\mathbb{Z}^n} c_k(f)\phi(x-k), \tag{7}$$

or in Fourier domain

$$\hat{f}(A^T\omega) = \frac{1}{|\det A|^{1/2}} c_f(\omega)\hat{\phi}(\omega), \tag{8}$$

in L^2 sense and with an appropriate $2\pi\mathbb{Z}^n$-periodic function $c_f(\omega)$ with Fourier coefficients $(c_k(f))_{k\in\mathbb{Z}^n}$. Then $c_f(\omega)$ can be decomposed with respect to the cosets:

$$\begin{aligned} c_f(\omega) &= \sum\nolimits_{k\in\mathbb{Z}^n} c_k(f) e^{i\langle\omega,k\rangle} = \sum\nolimits_{m=0}^{q} \sum\nolimits_{k\in\gamma_m + A\mathbb{Z}^n} c_k(f) e^{i\langle\omega,k\rangle} \\ &= \sum\nolimits_{m=0}^{q} e^{i\langle\omega,\gamma_m\rangle} \sum\nolimits_{k\in A\mathbb{Z}^n} c_{k+\gamma_m}(f) e^{i\langle\omega,k\rangle} = \sum\nolimits_{m=0}^{q} c_f^m(\omega), \end{aligned}$$

where

$$\begin{aligned} c_f^m(\omega) &= e^{i\langle\omega,\gamma_m\rangle} \sum\nolimits_{k\in A\mathbb{Z}^n} c_{k+\gamma_m}(f) e^{i\langle\omega,k\rangle} = e^{i\langle\omega,\gamma_m\rangle} \sum\nolimits_{k\in\mathbb{Z}^n} c_{Ak+\gamma_m}(f) e^{i\langle\omega,Ak\rangle} \\ &= e^{i\langle\omega,\gamma_m\rangle} \sum\nolimits_{k\in\mathbb{Z}^n} c_{Ak+\gamma_m}(f) e^{i\langle A^T\omega,k\rangle} = e^{i\langle\omega,\gamma_m\rangle} \kappa_f^m(A^T\omega). \end{aligned}$$

This representation exists for all functions V_1, in particular for ϕ and the wavelets. The following theorem indicates how the wavelets are constructed that generate the space W_0, such that $W_0 \oplus V_0 = V_1$.

Theorem 6. *Let $\phi \in V_0$ be a scaling function and let $\psi_1, \ldots, \psi_q \in V_1$. Then the family $\{\phi(\bullet - k)\}_{k\in\mathbb{Z}^n}$ is an orthonormal system if and only if*

$$\sum_{m=0}^{q} \left|\kappa_\phi^m(\omega)\right|^2 = 1 \quad \textit{almost everywhere.} \tag{9}$$

The system $\{\phi(\bullet - k)\}_{k\in\mathbb{Z}^n} \cup \bigcup_{m=1}^{q}\{\psi_m(\bullet - k)\}_{k\in\mathbb{Z}^n}$ *is an orthonormal basis in* V_1 *if and only if the so-called polyphase matrix*

$$\begin{pmatrix} \kappa_\phi^0(\omega) & \kappa_{\psi_1}^0(\omega) & \cdots & \kappa_{\psi_q}^0(\omega) \\ \vdots & \vdots & \ddots & \vdots \\ \kappa_\phi^q(\omega) & \kappa_{\psi_1}^q(\omega) & \cdots & \kappa_{\psi_q}^q(\omega) \end{pmatrix}$$

is unitary for almost all $\omega \in \mathbb{R}^n$.

The proof for a more general version of this theorem is given in [74, Sect. 5.2].

A summary and a condition for r-regular wavelets yields in the following theorem.

Theorem 7. *Consider a multiresolution analysis on* $\mathbb{R}^n$ *associated with a dilation matrix* A.

(i) Then there exists an associated wavelet set consisting of $q = |\det A| - 1$ *functions.*
(ii) If the multiresolution analysis is r*-regular and in addition* $2q + 1 > n$*, then there exists an associated wavelet set consisting of* q *functions, which all are* r*-regular.*

The idea of the proof is that for an r*-regular function* ϕ *on* $\mathbb{R}^n$ *and a* $2\pi\mathbb{Z}^n$*-periodic* C^∞*-function* $\eta(\omega)$*, the convolution* ψ *defined by* $\hat{\psi}(\omega) = \eta(\omega)\hat{\phi}(\omega)$ *is an* r*-regular function. For an explicit proof, see again [74].*

Example 4. As a continuation of Example 1, the wavelet function corresponding to $\phi = \chi_{[0,1)}$ is derived. To this end, consider the space $L^2(\mathbb{R})$ and the dilation $A = 2$. Then $q = \det A - 1 = 1$; thus $\gamma_0 = 0$ and $\gamma_1 = 1$ are representatives of the cosets of A. That is, there is only a single wavelet needed to generate W_0. Equation (7) yields

$$\frac{1}{\sqrt{2}}\phi\left(\frac{x}{2}\right) = \frac{1}{\sqrt{2}}\phi(x) + \frac{1}{\sqrt{2}}\phi(x-1)$$

for the normalized generator of V_1. Thus, $c_0(\phi) = \frac{1}{\sqrt{2}}$ and $c_1(\phi) = \frac{1}{\sqrt{2}}e^{i\omega}$. This implies $\kappa_\phi^0\left(\frac{\omega}{2}\right) = \kappa_\phi^1\left(\frac{\omega}{2}\right) = \frac{1}{\sqrt{2}}$. Then by Eq. (9) of Theorem 6, the family $\{\phi(\cdot - k)\}_{k\in\mathbb{Z}}$ is orthogonal, since $\kappa_\phi^0(\omega)^2 + \kappa_\phi^1(\omega)^2 = 1$. The polyphase matrix

$$\begin{pmatrix} \kappa_\phi^0(\omega) & \kappa_\psi^0(\omega) \\ \kappa_\phi^1(\omega) & \kappa_\psi^1(\omega) \end{pmatrix}$$

can be completed to a unitary matrix by choosing $\kappa_\psi^0(\omega) = \frac{1}{\sqrt{2}} = -\kappa_\psi^1(\omega)$. The corresponding wavelet then has the representation

$$\frac{1}{\sqrt{2}}\psi\left(\frac{x}{2}\right) = \frac{1}{\sqrt{2}}\phi(x) - \frac{1}{\sqrt{2}}\phi(x-1),$$

corresponding to (7). This yields the Haar wavelet ψ as illustrated in Fig. 2.

B-Splines

Several of the criteria for scaling functions and multiresolution analyses given in the previous section are based on the Fourier representation of the scaling function, e.g., the Riesz sequence criterion and the orthonormalization trick in Theorem 4, as well as the criterion for the order of approximation in Theorem 5. For this reason, the modeling of a scaling function in the Fourier domain to achieve certain specific properties is promising.

Aiming at constructing a scaling function $\phi \in L^2(\mathbb{R})$ of regularity $r = -1, 0, 1, \ldots$, this property is considered in the Fourier domain: It is a decay property of the Fourier transform $\hat{\phi}$ (compare with section "Regularity and Decay Under the Fourier Transform"):

$$\hat{\phi}(\omega) = \mathcal{O}\left(\frac{1}{\|\omega\|^{r+2}}\right) \quad \text{for } \|\omega\| \to \infty.$$

Taking into account Theorem 2, a first model for the scaling function in the Fourier domain is

$$\hat{\phi}(\omega) = \frac{\nu(\omega)}{\omega^{r+2}}, \quad \omega \in \mathbb{R}, \tag{10}$$

where the function ν still has to be specified. Since scaling functions satisfy a scaling relation (3)

$$\phi\left(\frac{x}{2}\right) = \sum\nolimits_{k\in\mathbb{Z}} h_k \phi(x-k) \quad \text{in } L^2(\mathbb{R}),$$

the Fourier transform of this equation yields

$$2\hat{\phi}(2\omega) = H(\omega)\hat{\phi}(\omega),$$

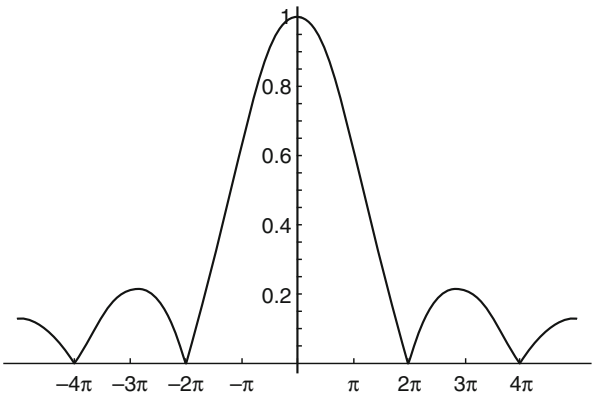


Fig. 4 The function $|\beta^0 \wedge |$ is strictly positive in the interval $[-\pi, \pi]$

where $(h_k)_{k\in\mathbb{Z}}$ is the sequence of Fourier coefficients of the 2π-periodic function H. For the ansatz (10),

$$H(\omega) = 2\frac{\widehat{\phi(2\omega)}}{\hat{\phi}(\omega)} = 2\frac{\nu(2\omega)}{(2\omega)^{r+2}}\frac{\omega^{r+2}}{\nu(\omega)} = \frac{1}{2^{r+1}}\frac{\nu(2\omega)}{\nu(\omega)}. \tag{11}$$

This gives the criteria for the choice of the function ν:

(i) ν vanishes at the origin and there is a zero of order $r + 2$. This ensures that $\hat{\phi} \in L^2(\mathbb{R})$ and that Theorem 3(iii) is satisfied.
(ii) $\nu(2\omega)$ is a trigonometric function, to ensure that $H(\omega)$, the so-called scaling filter, is 2π periodic.
(iii) ν has no other zeros in $[-\pi, \pi]$, except at the origin. Otherwise, the auto-correlation filter $A(\omega) = \sum_{k\in\mathbb{Z}} |\hat{\phi}(\omega + 2\pi k)|^2$ would vanish somewhere, and the shifts of the function ϕ would fail to generate a Riesz sequence; see Theorem 4(i).

A simple function ensuring all three requirements (i), (ii), and (iii) is

$$\nu(\omega) = (\sin(\omega/2)\theta(\omega/2))^{r+2},$$

where θ is a 2π-periodic phase factor such that $|\theta| = 1$, i.e., a shift in time domain. Choosing $\theta(\omega) = e^{-i\omega}$ yields the cardinal B-splines as given in (1) resp. (2):

$$\widehat{\beta^0}(\omega) = \int_0^1 e^{-i\omega t}\,dt = \frac{1 - e^{-i\omega}}{i\omega} = \frac{\sin(\omega/2)}{\omega/2}e^{-i\omega/2}.$$

Since β^0 has compact support, $\hat{\beta}^0 \in C^\infty$ [56]. Due to the convolution formula (2),

$$\hat{\beta}^m(\omega) = \left(\frac{1-e^{-i\omega}}{i\omega}\right)^{m+1} = \left(\frac{\sin(\omega/2)}{\omega/2}e^{-i\omega/2}\right)^{m+1}. \tag{12}$$

The β^m are scaling functions of regularity $r = m-1$, as the verification of the criteria in Theorem 3 shows. In fact, the following holds. Let $A = 2$.

Integrability: Since by (12) the functions $\hat{\beta}^m$ are L^2 integrable, so are the β^m, $m \in \mathbb{N}_0$.

Riesz sequence property: The shifted characteristic functions $\beta^0(x-k) = \chi_{[0,1)}(x-k)$, $k \in \mathbb{Z}$, are clearly orthonormal. Theorem 4(ii) thus yields

$$\sum\nolimits_{k\in\mathbb{Z}} |\hat{\beta}^0(\omega+2\pi k)|^2 = 1 \quad \text{almost everywhere.}$$

To verify the Riesz sequence property for β^m, the autocorrelation filter must be bounded with strictly positive constants from above and from below. It is

$$\sum\nolimits_{k\in\mathbb{Z}} |\hat{\beta}^m(\omega+2\pi k)|^2 = \sum\nolimits_{k\in\mathbb{Z}} |\hat{\beta}^0(\omega+2\pi k)|^{2m+2}.$$

In $[-\pi,\pi]$, $|\hat{\beta}^0|$ is clearly positive (cf. Fig. 4), which gives $|\hat{\beta}^0(\pi)| = 2/\pi$ as a positive bound from below. There is a constant c, such that

$$0 < c = (2/\pi)^{2m+2} < |\hat{\beta}^0(\omega)|^{2m+2} \le \sum\nolimits_{k\in\mathbb{Z}} |\hat{\beta}^m(\omega+2\pi k)|^{2m+2}.$$

Since the sequence $(|\hat{\beta}^0(\omega+2\pi k)|)_{k\in\mathbb{Z}} \in l^2(\mathbb{Z})$ for all $\omega \in \mathbb{R}$, the same is true for the sequence $|\hat{\beta}^m(\omega+2\pi k)| = |\hat{\beta}^0(\omega+2\pi k)|^{m+1}$. This yields the existence of the requested upper bound $c_2 < \infty$. Thus $\{\hat{\beta}^m(\bullet - k)\}_{k\in\mathbb{Z}}$ forms a Riesz sequence in $L^2(\mathbb{R})$.

Scaling relation: The scaling filter (11)

$$H(\omega) = 2^{-m}\frac{(1-e^{-i2\omega})^{m+1}}{(1-e^{-i\omega})^{m+1}} = 2^{-m}(1+e^{-i\omega})^{m+1} = 2^{-m}\sum\nolimits_{k=0}^{m+1}\binom{m+1}{k}e^{-i\omega k}$$

is obviously 2π-periodic and has Fourier coefficients $\left(2^{-m}\binom{m+1}{k}\right)_{k\in\mathbb{Z}} \in l^2(\mathbb{Z})$. Hence, the B-splines satisfy the scaling relation (3)

$$\beta^m(x/2) = \sum\nolimits_{k=0}^{m+1} 2^{-m}\binom{m+1}{k}\beta^m(x+k).$$

For β^0 this equation reads

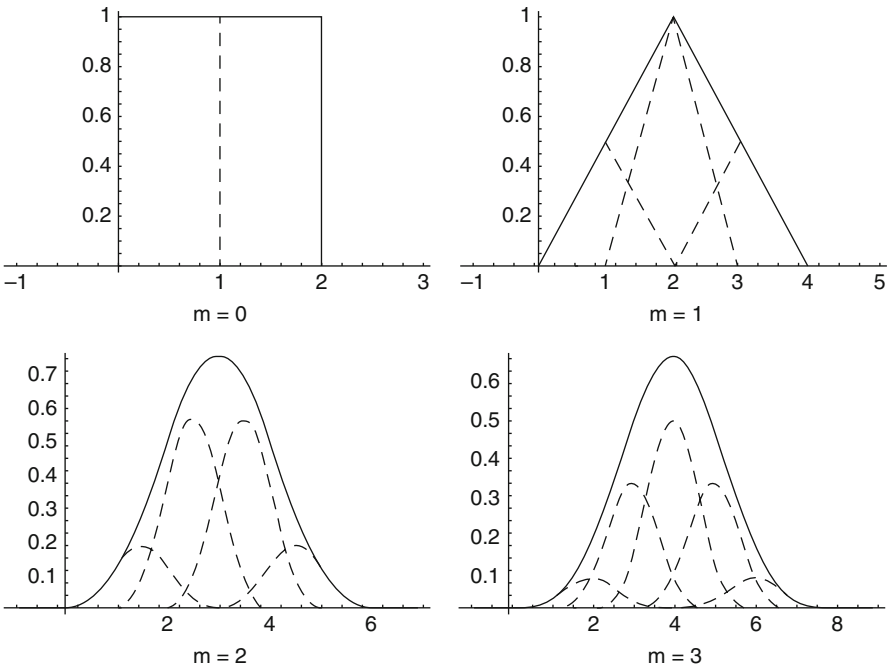


Fig. 5 Scaling relation for B-splines β^m, $m = 0, \ldots, 3$. The B-spline versions $\beta^m(x/2) \in V_{-1}$ are displayed with *solid lines*; the scaled translates in V_0 are depicted with *dashed lines*. The sum of the dashed functions gives the B-spline at the lower scale $\beta^m(x/2)$

$$\beta^0(x/2) = \beta^0(x) + \beta^0(x+1),$$

which is true since $\beta^0(x/2) = \chi_{[0,1)}(x/2) = \chi_{[0,2)}(x)$. This equation and examples for scaling relations of other B-splines are illustrated in Fig. 5.

Continuity and positivity of $\hat{\phi}$ at the origin: From Eq. (12),

$$|\hat{\beta}^m(\omega)| = \left| \frac{\sin(\omega/2)}{\omega/2} \right|^{m+1},$$

which has a continuous continuation at the origin, and $\hat{\beta}^m(0) = 1$. Thus we have proved the following conclusion:

Theorem 8. *The cardinal B-spline β^m, $m \in \mathbb{N}_0$, is a scaling function of an $m-1$-regular multiresolution analysis with dilation 2. The order of approximation is $m+1$.*

Note that the cardinal B-splines β^m with Fourier transform of the form (12) are scaling functions, but they are not yet orthonormalized: The family $\{\beta^m(\cdot - k)\}_{k\in\mathbb{Z}}$ spans V_0 and is a Riesz basis, but it is not an orthonormal basis of V_0. Orthonormality can be achieved with Theorem 8 and Eq. (4):

$$\hat{B}^m(\omega) := \frac{\hat{\beta}^m(\omega)}{\sqrt{\sum_{k\in\mathbb{Z}} |\hat{\beta}^m(\omega + 2\pi k)|^2}}.$$

Figure 6 shows some orthonormalized B-spline scaling functions and the corresponding wavelets.

Polyharmonic B-Splines

The same approach to model scaling functions in Fourier domain can be done in higher dimensions. We aim at constructing a scaling function for a multiresolution analysis of $L^2(\mathbb{R}^n)$ of the form

$$\hat{\phi}(\omega) = \frac{\nu(\omega)}{\| \omega \|^{2r}}, \quad r \in \mathbb{N}, \quad r > n/2, \quad x \in \mathbb{R}^n.$$

With an appropriate trigonometric polynomial

$$\nu(\omega) = \left(4 \sum_{k=1}^{n} \sin^2(\omega_k/2)\right)^r, \quad \omega = (\omega_1, \ldots, \omega_n),$$

$\hat{\phi}$ is a nonseparable scaling function for a multiresolution analysis of $L^2(\mathbb{R}^n)$ with respect to dilation matrices A that are scaled rotations. The corresponding function in space domain ϕ is called elementary r-harmonic cardinal B-spline, or short polyharmonic B-spline $\mathcal{P}^r$. This terminology can be justified as follows. The Fourier transform in the sense of tempered distributions of the function $1/\| \omega \|^{2r}$ is indeed a polynomial – up to a logarithmic factor for $2r - n$ even. In fact, in $\mathcal{S}'(\mathbb{R}^n)$,

$$\mathcal{F}^{-1}(1/\| \bullet \|^{2r})(x) = \| x \|^{2r-n} \, (A(n,r) \ln \| x \| + B(n,r)) =: \rho(x),$$

with constants $A(n,r)$, $B(n,r)$ as given in [63, Chap. VII, §7], and $A(n,r) = 0$ except for $2r - n$ even. (Note that for $r > n/2$ on the right-hand side, the Hadamard finite parts have to be considered.) The term polyharmonic comes from the fact that ρ is the Green function of the r-iterated Laplace operator Δ^r. However, with these considerations,

$$\phi(x) = \mathcal{P}^r(x) = \sum_{k\in\mathbb{Z}^2} \nu_k \rho(x + k) \quad \text{almost everywhere}$$

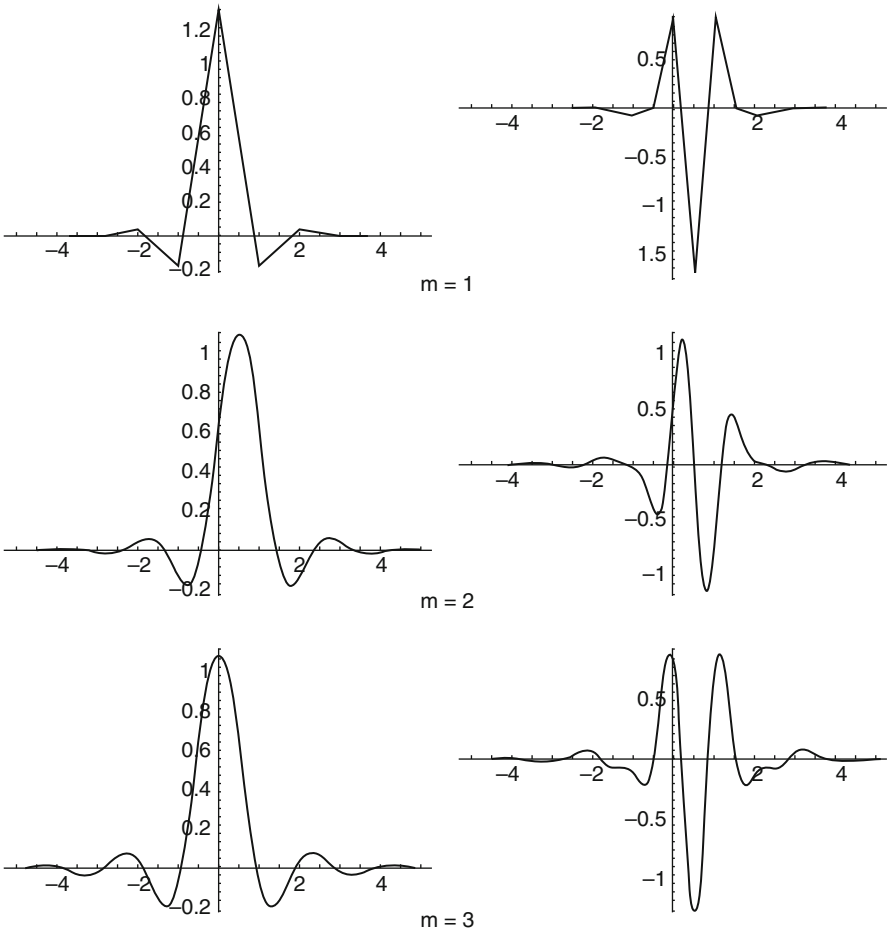


Fig. 6 Orthonormalized B-splines B^m (*left column*) and corresponding wavelets (*right column*) for $m = 1, 2, 3$ in time domain. Note that the orthonormalized B-splines and wavelets do not have compact support. Due to the orthonormalization procedure (4), the orthonormalized B-spline is an infinite series of shifted, compactly supported B-splines

becomes an nD spline. Here $(\nu_k)_{k\in\mathbb{Z}}$ is the sequence of the Fourier coefficients of ν. Due to the decay in Fourier domain, the polyharmonic B-spline $\mathcal{P}^r$ has continuous derivatives D^β for multi-indices $|\beta| < 2r - n$. In the same way as for the B-splines, it can be shown with the theorems given in section "Mathematical Foundations" that ϕ forms indeed a scaling function with approximation order $2r$ [53,54,71]. Figure 7 shows the polyharmonic B-spline scaling function in space domain and in frequency domain.

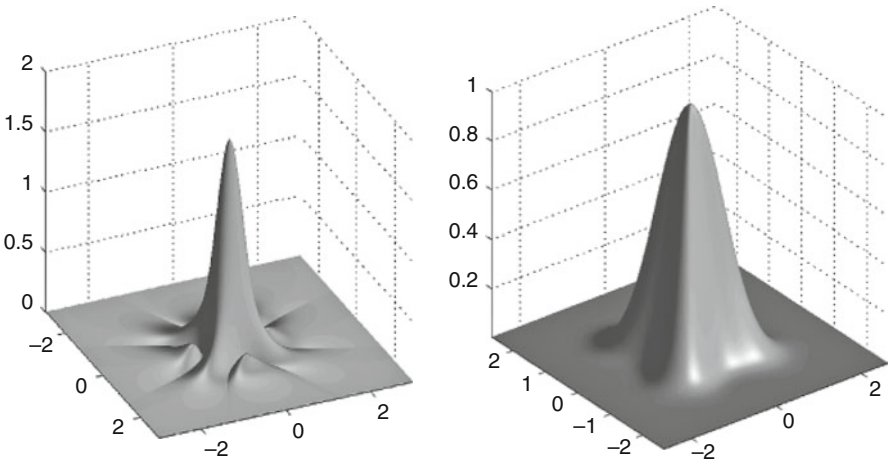


Fig. 7 Polyharmonic B-spline for $r = 3$ in space domain (*left*) and frequency domain (*right*)

4 Survey on Spline Families

There are many function families that consist of piecewise polynomials and that are called splines, which in addition fulfill a multiresolution condition in the one or the other sense. These families can be classified by various aspects, e.g., by their dimensionality, by the lattice which is invariant under the corresponding dilation matrix, by the geometries they are defined on, or whether they provide phase information or not, and so on. The following sections list some of these spline approaches and illustrates their mathematical properties and features.

Schoenberg's B-Splines for Image Analysis: The Tensor Product Approach

As mentioned in Example 3, multiresolution analyses for $L^2(\mathbb{R}^n)$ and dilation matrix $2Id$ can be generated from tensor products of 1D dyadic multiresolution analyses. To analyze images with B-splines, the tensor product $\beta^m(x)\beta^m(y)$ of B-splines is a scaling function for $L^2(\mathbb{R}^2)$ and the dilation matrix $A = 2Id$. Since in 2D the determinant $\det A = 4$, the corresponding detail space W_0 is spanned by three wavelet functions:

$$\psi(x)\beta^m(y), \quad \beta^m(x)\psi(y), \quad \psi(x)\psi(y), \quad x, y \in \mathbb{R}. \tag{13}$$

A drawback of this approach is the fact that these wavelets prefer horizontal, vertical, and diagonal directional features and are not sensitive to other directions; see Fig. 8. For the analysis of images with many isotropic features, the use of

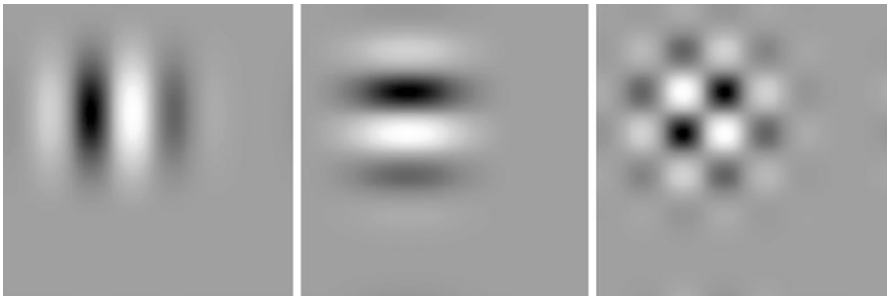


Fig. 8 The three B-spline tensor wavelets (13) show a preference for horizontal, vertical, and diagonal directions. Here, $m = 2$. Minimal resp. maximal function values are given in *black* resp. *white*

isotropic or steerable wavelets is recommended. However, the tensor approach is a simple and widely used wavelet approach. For an illustration of the respective image decomposition in coarse approximations and details of various sizes, see Fig. 9.

Fractional and Complex B-Splines

The B-splines as described up to now have a discrete order of smoothness, i.e., they are C^n functions with $n \in \{-1, 0, 1, 2, \ldots\}$. For some applications, e.g., in medical imaging, where the order of smoothness of certain image classes is fractional, it would be favorable to have a spline and wavelet basis that is adaptable with respect to this regularity [1, 37, 73]. A first step in this direction was done by T. Blu and M. Unser, who proposed B-splines and wavelets of fractional orders [5]. They defined two variants of fractional B-splines, the causal ones and the symmetrical ones.

The causal fractional B-spline is generated by applying the $(\alpha + 1)$ fractional difference operator to the one-sided power function ${t_+}^\alpha$:

$$\beta_+^\alpha(t) := \frac{1}{\Gamma(\alpha+1)} \Delta_+^{\alpha+1} t_+^\alpha = \frac{1}{\Gamma(\alpha+1)} \sum\nolimits_{k \geq 0} (-1)^k \binom{\alpha+1}{k} (t-k)_+^\alpha.$$

The Fourier-transform representation is similar to the one of the classical B-splines (cf. Eq. 12):

$$\hat{\beta}_+^\alpha(\omega) = \left(\frac{1 - e^{-i\omega}}{i\omega} \right)^{\alpha+1}.$$

Here again, the smoothness property $\beta_+^\alpha \in C^{m,\gamma}(\mathbb{R})$ in the time domain is gained by the fractional polynomial decay of order $\mathcal{O}(|\omega|^{\alpha+1})$ in the frequency domain. Note that $C^{m,\gamma}(\mathbb{R})$ denotes the Hölder space with exponent $m = \lfloor \alpha + 1 \rfloor$ and

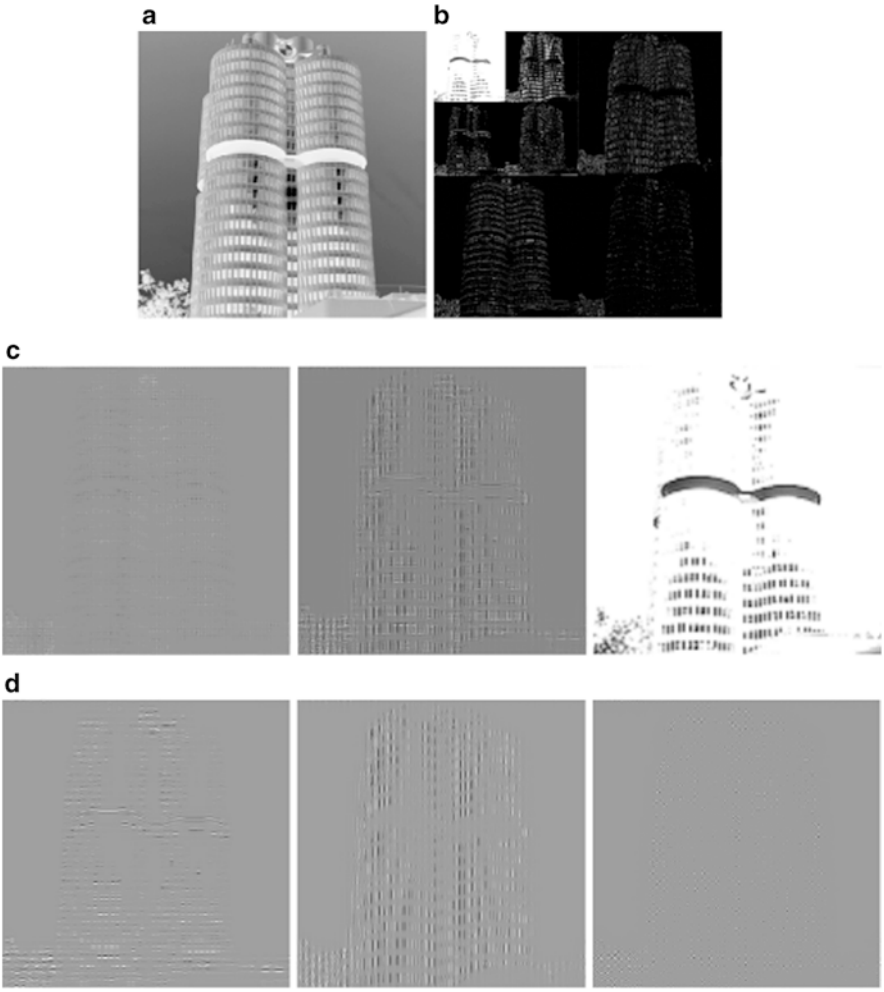


Fig. 9 Decomposition of an image [18, Part of IM008.tif] into coarse approximations and details of various sizes. (**a**) Original image. (**b**) Matrix of the absolute values of the multiresolution coefficients. Large coefficients are *white*. The wavelet coefficients are depicted in the lower two and the upper right band of each scale and the approximation coefficients in the upper left band. (**c**) Two steps of the multiresolution decomposition. From *left to right*: finest details and second finest details, remaining approximation of the original image. (**d**) Second finest details (*c, center*) split into the contribution of the three wavelets. From *left to right*: the decomposition into horizontal, vertical, and diagonal details

$\gamma = \alpha + 1 - m$, i.e., the space of m-times continuously differentiable functions f that are Hölder regular with exponent $0 < \gamma \leq 1$ such that there is a constant $C > 0$ with

$$|D^m f(t) - D^m f(s)| \leq C|t-s|^\gamma \quad \forall s,t \in \mathbb{R}.$$

Although the fractional B-splines are not compactly supported, they decay in the order $\mathcal{O}(|t|^{-(\alpha+2)})$ as $t \to \infty$. They are elements of $L^1(\mathbb{R})$ for $\alpha > 0$, of $L^2(\mathbb{R})$ for $\alpha > -\frac{1}{2}$, and of the Sobolev spaces $W_2^r(\mathbb{R})$ for $r < \alpha + \frac{1}{2}$. They share many properties with their classical B-spline relatives, such as the convolution property, and their relation to difference operators, i.e., they are integral kernels for fractional difference operators [26] and they are scaling functions for dyadic multiresolution analyses. This can be verified by the procedure given in the section "B-Splines".

The causal fractional B-spline is not symmetric. Since for some signal and image analysis tasks symmetrical bases are preferred, in [5] the symmetrical fractional B-splines β_*^α are proposed. They are defined in Fourier domain as follows:

$$\hat{\beta}_*^\alpha(\omega) = \left(\frac{1-e^{-i\omega}}{i\omega}\right)^{\frac{\alpha+1}{2}} \left(\frac{1+e^{i\omega}}{-i\omega}\right)^{\frac{\alpha+1}{2}} = \left|\frac{\sin(\omega/2)}{\omega/2}\right|^{\alpha+1}, \tag{14}$$

and therefore obviously are symmetrical in the time domain. The same regularity and decay properties apply as for the causal fractional B-splines. The symmetrical fractional B-splines are also piecewise polynomials, as long as $\alpha \notin 2\mathbb{N}_0$. For even integer degrees, the singularity introduced through the absolute value in Eq. (14) causes that β_*^{2m} is a sum of integer shifts of the logarithmic term $|t|^{2m}\ln(t)$ for $m \in \mathbb{N}_0$. For the explicit time-domain representation and further details on these splines, cf. [5].

In [6], Blu and Unser defined another variant, the generalized fractional B-spline or (α, τ)-fractional spline β_τ^α with a parameter $\tau \in \mathbb{R}$. Also, these splines are defined via their Fourier domain representation:

$$\hat{\beta}_\tau^\alpha(\omega) = \left(\frac{1-e^{-i\omega}}{i\omega}\right)^{\frac{\alpha+1}{2}+\tau} \left(\frac{1-e^{i\omega}}{-i\omega}\right)^{\frac{\alpha+1}{2}-\tau}.$$

As above, the parameter $\alpha > 0$ controls the regularity of the splines. The parameter τ, in contrast, controls the position of the splines with respect to the grid $2\mathbb{Z}$. This can be justified by the following fact. All variants of the B-splines considered in this section converge to the optimally time-frequency-localized functions in the sense of Heisenberg, i.e., to Gaussians or Gabor functions, if the degree α becomes large. For a proof for the classical cardinal B-splines, see [67]. In the case of the (α, τ)-fractional splines [6],

$$\beta_\tau^\alpha(t) = \mathcal{O}\left(e^{-\frac{6}{\alpha+1}(t-\tau)^2}\right) \quad \text{for } \alpha \to \infty.$$

This explains the notion "shift parameter" for τ. Moreover, the parameter τ allows to interpolate the spline family between the two "knots," the symmetrical ones ($\tau = 0$) and the causal ones $\left(\tau = \frac{\alpha+1}{2}\right)$; see Fig. 10. Both parameters α and τ can be tuned

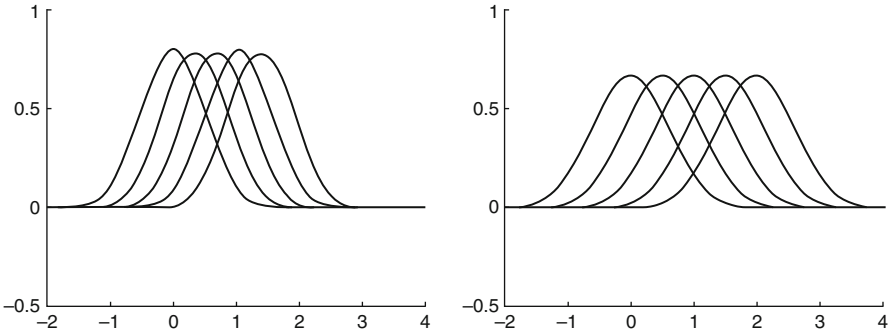


Fig. 10 The fractional (α, τ)-splines interpolate the families of the causal and the symmetric fractional splines. $\tau = \frac{\alpha+1}{2}k$ for $k = 0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1$ from the most *right* (causal) to the most *left* (symmetrical) function in each image. *Right*: $\alpha = 1.8$. *Left*: $\alpha = 3$

independently and therefore allow for an individual adjustment of the analyzing basis.

Another generalization are the complex B-splines [27]. There are two variants, both defined via their Fourier domain representation. Let $z = \alpha + i\gamma \in \mathbb{C}, \alpha > -\frac{1}{2}, \gamma \in \mathbb{R}$, and $y \in \mathbb{C}$. Then

$$\hat{\beta}^z(\omega) = \left(\tfrac{1-e^{-i\omega}}{i\omega}\right)^{z+1},$$
$$\hat{\beta}^z_y(\omega) = \left(\tfrac{1-e^{-i\omega}}{i\omega}\right)^{\frac{z+1}{2}-y} \left(\tfrac{1-e^{i\omega}}{-i\omega}\right)^{\frac{z+1}{2}+y}$$

are complex B-splines of complex degree z. The functions are well defined, because the function $\Omega(\omega) = \frac{1-e^{-i\omega}}{i\omega}$ never touches the negative real axis such that $\Omega(\omega)^z$ is uniquely defined. β^z and β^z_y are elements of the Sobolev spaces $W_2^r(\mathbb{R})$ for $r < \alpha + \frac{1}{2}$. β^z has the time-domain representation

$$\beta^z(t) = \frac{1}{\Gamma(z+1)} \sum\nolimits_{k\geq 0} (-1)^k \binom{z+1}{k} (t-k)^z_+,$$

i.e., β^z is a piecewise polynomial of complex degree. For more details on the properties of these families of complex splines, cf. [27].

The idea behind the complex degree is as follows: The real part $\mathrm{Re}\, z = \alpha$ operates as regularity and decay parameter in the same way as for the fractional B-splines. The imaginary part, however, causes an enhancement resp. damping of positive or negative frequencies. In fact,

$$\hat{\beta}^z(\omega) = \hat{\beta}^\alpha_+(\omega) e^{-i\gamma \ln|\Omega(\omega)|} e^{\gamma \arg \Omega(\omega)}.$$

The imaginary part γ of the complex degree introduces a phase and a scaling factor in the frequency domain. The frequency components on the negative and positive

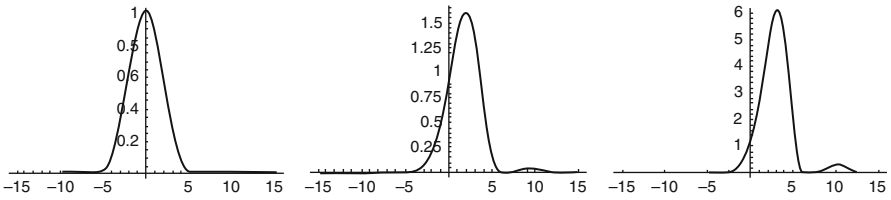


Fig. 11 The frequency spectrum $|\beta^{z\wedge}|$ for $z = 3 + i\gamma$, $\gamma = 0, 1, 2$ (from *left to right*). The spectrum of $\beta^3 = \beta_+{}^3$ is symmetric (*right*), whereas the spectra of β^{3+i} (*center*) and β^{3+2i} (*left*) show an enhancement of the positive frequency axis

real axis are enhanced with different signs, because $\arg \Omega(\omega) \geq 0$ for negative ω and $\arg \Omega(\omega) \leq 0$ for positive ω. Figure 11 illustrates this effect.

With real-valued functions, only symmetric spectra can be analyzed. The complex B-splines, however, allow for an approximate analysis of the positive or the negative frequencies, because the respective symmetric bands are damped due to the complex exponent. However, the complex B-splines inherit many properties of their classical and fractional relatives.

All of the generalized B-spline families mentioned in this section have in common that they are scaling functions of dyadic multiresolution analyses. They are one-dimensional functions, but with the tensor approach mentioned in Example 3 and the section "Schoenberg's B-Splines for Image Analysis – the Tensor Product Approach", they are also suitable for image processing tasks. Although the fractional and the complex splines, in general, do not have compact support, they allow for fast analysis and synthesis algorithms. Due to their closed form in Fourier domain, they invite for an implementation of these algorithms in Fourier domain.

Polyharmonic B-Splines and Variants

In the section "Polyharmonic B-Splines", the so-called polyharmonic B-splines in $\mathbb{R}^n$ were introduced. They are defined in the Fourier domain by the representation

$$\hat{\mathcal{P}}^r(\omega) = \left(\frac{4 \sum_{k=1}^n \sin^2(\omega_k/2)}{\sum_{k=1}^n \omega_k^2} \right)^r, \quad r > n/2, \ \omega = (\omega_1, \ldots, \omega_n).$$

These polyharmonic B-splines satisfy many properties of the classical Schoenberg splines; e.g., they are piecewise polynomial functions, they satisfy a convolution relation $\mathcal{P}^{r_2+r_2} = \mathcal{P}^{r_1} * \mathcal{P}^{r_2}$, they are positive functions, etc. However, they do not share the property that they converge to the optimally space-frequency-localized Gaussians as r increases [71]. This is due to the fact that the trigonometric polynomial in the numerator regularizes insufficiently at the origin: The second-order derivative of

$$\frac{4\sum_{k=1}^{n}\sin^2(\omega_k/2)}{\sum_{k=1}^{n}\omega_k^2}$$

is not continuous. Van De Ville et al. [71] therefore proposed another localizing trigonometric polynomial:

$$\mu(\omega) = 4\sum_{k=1}^{n}\sin^2(\omega_k/2) - \frac{8}{3}\sum_{k=1}^{n-1}\sum_{l=k+1}^{n}\sin^2(\omega_k/2)\sin^2(\omega_l/2). \tag{15}$$

A new function family then is defined in the Fourier domain via

$$\hat{\mathcal{Q}}^r(\omega) = \left(\frac{\mu(\omega)}{\|\omega\|^2}\right)^r,\ r > \frac{n}{2}. \tag{16}$$

$\mathcal{Q}^r$ is called isotropic polyharmonic B-spline. The function is piecewise polynomial (except for $2r–n$ even, where a logarithmic factor has to be added; see below) and shares with the polyharmonic splines their decay properties in Fourier domain and their regularity properties in space domain. $\mathcal{Q}^r$ converges to a Gaussian as r increases, which makes the function family better suitable for image analysis than $\mathcal{P}^r$, because of the better space-frequency localization. This effect is due to the higher-order rotation invariance or isotropy of the localizing trigonometric polynomial (15):

$$\nu(\omega) = 1 + \mathcal{O}(\|\omega\|^2) \text{ vs. } \mu(\omega) = 1 + \frac{1}{12}\|\omega\|^2 + \mathcal{O}(\|\omega\|^4) \quad \text{as } \|\omega\| \to 0.$$

This causes that $\hat{\mathcal{Q}}^r$ has a second-order moment, and thus the central limit theorem can be applied to proof the convergence to the Gaussian function. In addition, $\hat{\mathcal{Q}}^r$ has a higher regularity than $\hat{\mathcal{P}}^r$; therefore, $\mathcal{Q}^r$ decays faster. For a complete proof of the localization property, see [71]. An example of the isotropic polyharmonic spline is given in Fig. 12.

The polyharmonic B-splines and the isotropic polyharmonic B-splines both are real-valued functions. The isotropic B-spline is approximately rotation invariant and therefore is suited for image analysis of isotropic features. A complex-valued variant of these B-splines in 2D was introduced in [28]. The idea is to design a spline scaling function that is approximately rotation covariant, instead of rotation invariant. Rotation covariant here means that the function intertwines with rotations up to a phase factor.

Again, the design of the scaling function is done in the Fourier domain, now using a perfectly rotation-covariant function

$$\hat{\rho}_{r,N}(\omega_1,\omega_2) = \frac{1}{(\omega_1^2+\omega_2^2)^r(\omega_1 - i\,\omega_2)^N},$$

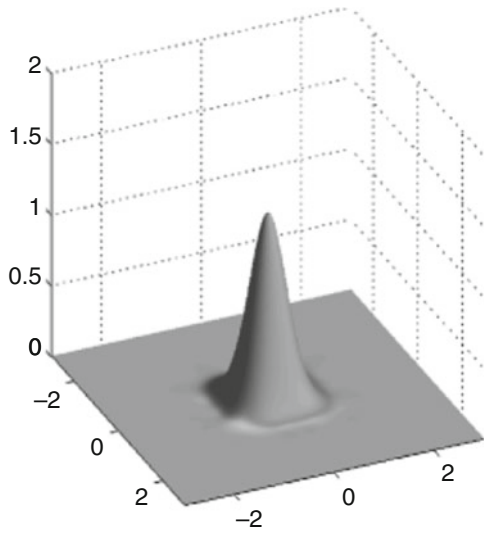


Fig. 12 The isotropic polyharmonic spline $\mathcal{Q}^3$. Compare with Fig. 7 of the classical polyharmonic B-spline $\mathcal{P}^3$

where $r \geq 0$ and $N \in \mathbb{N}$. In fact, for some rotation matrix $R_\theta \in GL(2,\mathbb{R})$, $\hat{\rho}_{r,N}(R_\theta\omega) = e^{-iN\theta}\hat{\rho}_{r,N}(\omega)$. For localizing the function $\hat{\rho}_{r,N}$, the same trigonometric polynomials ν and μ as above can be used. The corresponding complex polyharmonic B-splines are then defined in the frequency domain as

$$\begin{aligned} \hat{\mathcal{R}}_\nu^{r,N}(\omega_1,\omega_2) &= \frac{(\nu(\omega_1,\omega_2))^{r+\frac{N}{2}}}{\left(\omega_1^2+\omega_2^2\right)^r(\omega_1-i\omega_2)^N}, \quad \text{or as} \\ \hat{\mathcal{R}}_\mu^{r,N}(\omega_1,\omega_2) &= \frac{(\mu(\omega_1,\omega_2))^{r+\frac{N}{2}}}{\left(\omega_1^2+\omega_2^2\right)^r(\omega_1-i\omega_2)^N}. \end{aligned} \tag{17}$$

The case $N = 0$ yields the real-valued polyharmonic splines.

There are also other trigonometric polynomials that are suitable as localizing numerators for the real and the complex polyharmonic B-splines. With an appropriate choice, the features of the polyharmonic splines can be tuned [28]. However, for both the real and the complex variant, the localizing multiplier has to fulfill moderate conditions to make the respective polyharmonic B-splines become a scaling function. In 2D, the following result holds (cf. [28]):

Theorem 9. *Let $r > 0$ and $N \in \mathbb{N}_0$. Let $\eta(\omega_1, \omega_2)$ be a bounded, $2\pi\mathbb{Z}^2$-periodic function, such that*

$$\left| \frac{(\eta(\omega_1,\omega_2))^{r+\frac{N}{2}}}{\left(\omega_1^2+\omega_2^2\right)^r(\omega_1-i\omega_2)^N} \right|$$

is bounded in a neighborhood of the origin and such that $\eta(\omega_1, \omega_2) \neq 0$ for all $(\omega_1,\omega_2) \in [-\pi,\pi]^2\backslash\{(0,0)\}$.

Then $\hat{\phi} = \eta^{r+\frac{N}{2}} \cdot \hat{\rho}$ is the Fourier transform of a scaling function ϕ which generates a multiresolution analysis $\ldots V_{-1} \subset V_0 \subset V_1 \ldots$ of $L^2(\mathbb{R}^2)$ with dilation matrix $A = \begin{pmatrix} a & b \\ -b & a \end{pmatrix}$, $a, b \in \mathbb{Z}$:

$$V_j = \overline{\text{span}\,\{|\det A|^{j/2}\phi(A^j \bullet - k), k \in \mathbb{Z}^2\}}.$$

From the Fourier domain representation immediately follows that $\hat{\phi} \in L^2(\mathbb{R}^2)$ for $r + \frac{N}{2} > \frac{1}{2}$ and that $\hat{\phi}$ decays as $|\hat{\phi}(\omega_1, \omega_2)| = \mathcal{O}(\|(\omega_1, \omega_2)\|^{-2r-N})$ when $\|\ (\omega_1, \omega_2)\ \| \to \infty$.

As a result, for all three variants of the polyharmonic B-splines, the classical ones $\mathcal{P}^r$, the isotropic ones $\mathcal{Q}^r$, and the complex ones $\mathcal{R}^{r,N}$, the following properties hold: They are scaling functions for multiresolution analysis. Their smoothness parameter r can be chosen fractional and must fulfill $r + \frac{N}{2} > n/2$ for integrability reasons. Then the scaling function in space domain is an element of the Sobolov space $\phi \in W_2^s(\mathbb{R}^2)$ for all $s < 2r + N - 1$. The explicit space domain representation is

$$\phi(x) = \sum\nolimits_{k \in \mathbb{Z}^2} \eta_k \rho_{r,N}(x + k)$$

for almost all $x \in \mathbb{R}^2$. Here, $(\eta_k)_{k \in \mathbb{Z}^2}$ denotes the Fourier coefficients of $\eta^{r+\frac{N}{2}}$. $\rho_{r,N}$ is the inverse Fourier transform of the Hadamard finite part $Pf(\hat{\rho}_{r,N}) \in \mathcal{S}'(\mathbb{R}^2)$. In fact, for $r \notin \mathbb{N}_0$,

$$\rho_{r,N}(x_1, x_2) = c_1 \left(x_1^2 + x_2^2\right)^{r-1} (x_1 + i x_2)^N$$

and for $r \in \mathbb{N}_0$,

$$\rho_{r,N}(x_1, x_2) = c_2 \left(x_1^2 + x_2^2\right)^{r-1} (x_1 + i x_2)^N \left(\ln \pi \sqrt{\ln\left(\pi x_1^2 + x_2^2\right)} + c_3\right)$$

with appropriate constants $c_1, c_2, c_3 \in \mathbb{C}$. This justifies the notion spline for the function families. They all have a closed form in the frequency domain. As in the case of the 1D cardinal B-splines, there is a fast analysis and synthesis algorithm using the frequency domain representation and the fast Fourier transform; cf. Sect. 5.

Splines on Other Lattices

Splines on the Quincunx Lattice

The tensor product of two 1D dyadic multiresolution analyses yields a 2D multiresolution analysis with dilation matrix $A = 2Id$; cf. Example 3. As a consequence, the scaling factor while moving from one approximation space V_0 to the next coarser space V_1 is $|\det A| = 4$. For some image processing applications, especially in

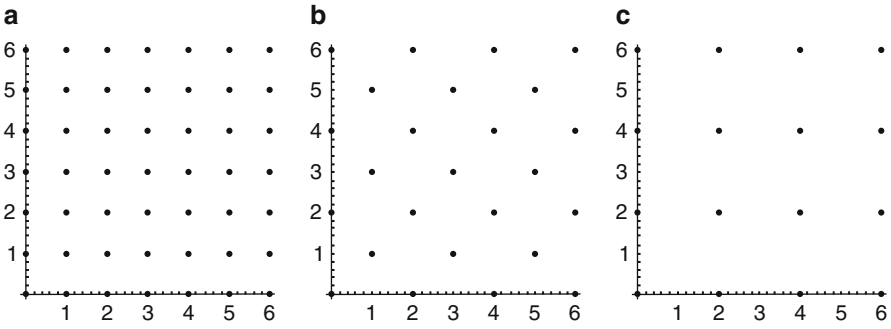


Fig. 13 Three iterations of the quincunx subsampling scheme. (**a**) $\mathbb{Z}^2$, (**b**) $A_q\mathbb{Z}^2$, (**c**) $A_q^2\ \mathbb{Z}^2$. The thinning of the $\mathbb{Z}^2$ lattice using the dilation matrix A_q is finer than dilation with the dyadic matrix $A = 2Id$, which in one step leads from (**a**) to (**c**)

medical imaging, this scaling step size is too large. A step size of 2 as in the 1D case would be preferred. Moreover, the decomposition of the wavelet space into three subspaces then would be avoided, and the eventual problematic of the directionality of the three involved wavelets would not arise. An example of a dilation matrix satisfying these requirement is the scaled rotation matrix

$$A_q = \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}$$

with $\det A_q = 2$. It leads to the so-called quincunx lattice. This lattice is generated by applying A_q to the Cartesian lattice. It holds $A_q\mathbb{Z}^2 \subset \mathbb{Z}^2$; see Fig. 13.

Since A_q falls into the class of scaled rotations, the polyharmonic B-spline construction including all variants is applicable for this case; cf. Theorem 9. Note that the tensor product approach in general is not suitable for the quincunx subsampling scheme.

Splines on the Hexagonal Lattice

Images as 2D objects are normally processed on the Cartesian lattice, i.e., the image pixels are arranged on a rectangular grid. For image processing, this arrangement has the drawback that not all neighbors of a pixel have the same relation: The centers of the diagonal neighbors have a larger distance to the center pixel than the adjacent ones. A higher degree of symmetry has the hexagonal lattice. It is therefore ideal for isotropic feature representation. The hexagonal lattice gives an optimal tessellation of $\mathbb{R}^2$ in the sense of the classical honeycomb conjecture, which says that any partition of the plane into regions of equal area has a perimeter at least that of regular hexagonal tiling [32]. The better isotropy of the hexagonal lattice is attractive for image analysis and has led to a series of articles on image processing methods (e.g., [31, 45, 52]) as well as on applications (e.g., [30, 36, 47, 72]).

The hexagonal lattice is generated by applying the matrix

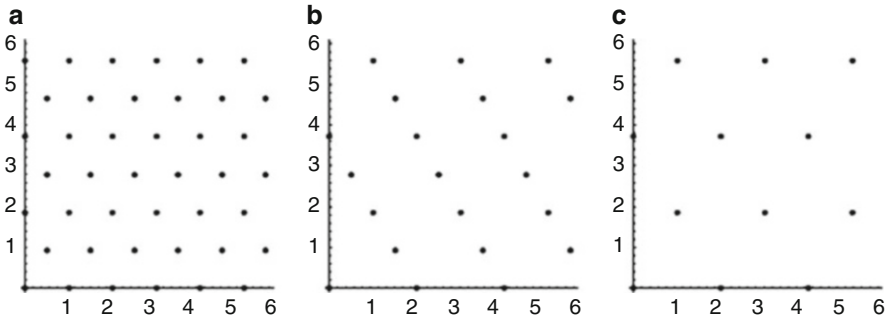


Fig. 14 Three iterations of subsampling of the hexagonal grid with $A = R_h B R_h^{-1}$, where $B = \left(\begin{smallmatrix} 1 & 1 \\ -1 & 1 \end{smallmatrix}\right)$. (**a**) Λ_h, (**b**) $A\Lambda_h = R_h B R_h^{-1} \Lambda_h = R_h B \mathbb{Z}^2$, (**c**) $A^2 \Lambda_h$

$$R_h = \sqrt{\frac{2}{\sqrt{3}}} \begin{pmatrix} 1 & 1/2 \\ 0 & \sqrt{3}/2 \end{pmatrix}$$

on the Cartesian lattice $\Lambda_h = R_h \mathbb{Z}^2$. A scaling function of a multiresolution analysis of $L^2(\mathbb{R}^2)$ in the hexagonal lattice fulfills all properties of Definition 1, but the last two. They change to

(iv) $f \in V_0 \Leftrightarrow f(\cdot - R_h k) \in V_0$ *for all* $k \in \mathbb{Z}^2$.
(v) *There exists a scaling function* $\phi \in V_0$, *such that the family* $\{\phi(\bullet - R_h k)\}_{k \in \mathbb{Z}^2}$ *of translates of* ϕ *forms a Riesz basis of* V_0.

Let A be a dilation matrix which leaves the hexagonal lattice invariant $A\Lambda_h \subset \Lambda_h$. Then A is of the form [19]

$$A = R_h B R_h^{-1}$$

with $B \in GL(2, \mathbb{R})$ having only integer entries and with eigenvalues strictly larger than one. Figure 14 gives an example of two subsampling steps on the hexagonal lattice.

There are several possible approaches to define spline functions on the hexagonal lattice. Sablonnière and Sbibih [57] proposed to convolve piecewise linear pyramids to generate higher-order B-splines. Van De Ville et al. [70] started with the characteristic function of one hexagon and also used an iterative convolution procedure to construct B-splines of higher degree. However, both approaches lead to discrete-order hexagonal B-splines. If A is a scaled rotation, then fractional and complex B-splines on the hexagonal lattice can be defined in an analog way as in the section "Polyharmonic B-Splines and Variants" for polyharmonic splines and their (complex) variants [19]. Consider again the perfectly rotation-covariant (or for $N = 0$ rotation-invariant) function

$$\hat{\rho}_{r,N}(\omega_1, \omega_2) = \frac{1}{\left(\omega_1^2 + \omega_2^2\right)^r (\omega_1 - i\omega_2)^N},$$

where $r > 0$ and $N \in \mathbb{N}_0$. The idea is now to use a hexagonal-periodic trigonometric polynomial for localizing this function and to eliminate the singularity at the origin. Condat et al. [19] proposed

$$\begin{aligned}\eta_h(\omega_1, \omega_2) = \frac{1}{\sqrt{3}}(6 - 2\,(\cos\,(3^{1/4}\,(-\omega_1/\sqrt{3} + \omega_2)/\sqrt{2}) \\ + \cos(3^{1/4}\,(\omega_1/\sqrt{3} + \omega_2)/\sqrt{2}) + \cos\,(3^{-1/4}\sqrt{2}\omega_1))),\end{aligned}$$

and defined the elementary polyharmonic hexagonal rotation-covariant B-spline via its frequency domain representation as

$$\hat{\mathcal{R}}_h^{r,N}(\omega_1, \omega_2) = \frac{(\eta_h(\omega_1, \omega_2))^{r+\frac{N}{2}}}{\left(\omega_1^2 + \omega_2^2\right)^r (\omega_1 - i\omega_2)^N}.$$

The B-spline in space domain then has the representation

$$\mathcal{R}_h^{r,N}(x) = \sum\nolimits_{k \in \mathbb{Z}^2} \eta_{h,k}\, \rho_{r,N}(x - R_h k).$$

Here, $(\eta_{h,k})_{k \in \mathbb{Z}^2}$ denotes the sequence of Fourier coefficients of $\eta_h^{r+\frac{N}{2}}$. Figure 15 shows the localizing trigonometric polynomial η and the Fourier spectra of two hexagonal splines.

For $N = 0$, the functions are the elementary polyharmonic hexagonal rotation-invariant B-splines. They are real-valued functions that converge to a Gaussian, as $r \to \infty$, and therefore are well localized in the space domain as well as in the frequency domain. For $N \in \mathbb{N}$, the splines are complex-valued functions and approximately rotation covariant in a neighborhood of the origin:

$$\hat{\mathcal{R}}_h^{r,N}(\omega) = e^{i N \arg(\omega)} \left(1 + C\,\|\omega\|^2 + \mathcal{O}(\|\omega\|^4)\right) \quad \text{for } \omega \to 0,$$

where $\omega = (\omega_1, \omega_2)$ and $C \in \mathbb{R}$ is a constant.

The translates of the complex B-spline $\mathcal{R}_h^{r,N}$ form a Riesz basis of the approximation spaces

$$V_j = \overline{\text{span}\left\{\mathcal{R}_h^{r,N}(A^j x - R_h k) : k \in \mathbb{Z}^2\right\}}^{L^2(\mathbb{R}^2)}, \quad j \in \mathbb{Z}.$$

The ladder of spaces $(V_j)_{j \in \mathbb{Z}}$ generates a multiresolution analysis of $L^2(\mathbb{R}^2)$ for the hexagonal grid and for scaled rotations A. Also in this case, the implementation of

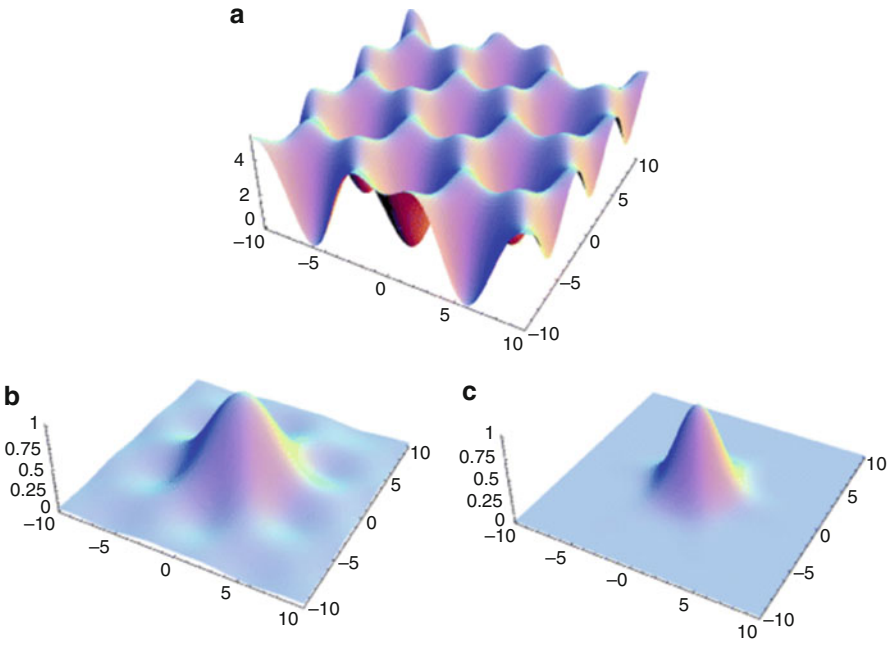


Fig. 15 Localization of $\rho^{\wedge}$ with (**a**) the hexagonal-periodic trigonometric polynomial η_h yields the elementary polyharmonic hexagonal rotation-covariant B-spline. Frequency spectrum of $\mathcal{R}_h^{\wedge} r, N$ (or equivalently of $\mathcal{R}_h^{\wedge} r + N/2, 0$), (**b**) for $r + \frac{N}{2} = 1$, and (**c**) for $r + \frac{N}{2} = 2.5$

the analysis and synthesis algorithm can be elegantly performed in the frequency domain [19, 71].

5 Numerical Implementation

For the illustration of the numerical method, we focus on the quincunx dilation matrix

$$A = \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix} \tag{18}$$

and consider the polyharmonic spline variants in 2D as defined in the section "Polyharmonic B-Splines and Variants". Since $\det A = 2$, the generators of the multiresolution space and the corresponding wavelet space are two functions: the scaling function (here the variant of the polyharmonic spline) and the associated wavelet. We consider the scaling function $\phi(x) = \mathcal{Q}_{\mu}^{r}(x)$ resp. $\phi(x) = \mathcal{R}_{\mu}^{r,N}(x)$. It spans the ladder of nested approximation spaces $\{V_j\}_{j \in \mathbb{Z}}$ via

$$V_j = \overline{\text{span}\,\{|\det A|^{j/2}\phi(A^j \bullet -k), k \in \mathbb{Z}^2\}}^{L^2(\mathbb{R}^2)}, \quad j \in \mathbb{Z}.$$

Denote

$$M(\omega) := \sum\nolimits_{k\in\mathbb{Z}^2} |\hat{\phi}(\omega + 2\pi k)|^2 \tag{19}$$

the autocorrelation filter. It is bounded $0 < M(\omega) \leq C$ for some positive constant C [28]. The scaling functions can be orthonormalized, applying the procedure given in Theorem 4(iii):

$$\hat{\phi}(\omega) = \frac{\hat{\phi}(\omega)}{\sqrt{M(\omega)}}.$$

The scaled shifts of Φ span the same spaces $\{V_j\}_{j\in\mathbb{Z}}$.

The B-splines as scaling functions ϕ satisfy a refinement relation

$$\phi(A^{-1}x) = \sum\nolimits_{k\in\mathbb{Z}^2} h_k \phi(x-k) \quad \text{almost everywhere and in } L^2(\mathbb{R}^2).$$

This relation in fact is a discrete convolution. The Fourier transform yields a $2\pi\mathbb{Z}^2$-periodic function $H \in L^2(\mathbb{T}^2)$ of the form

$$\begin{aligned} H(e^{i\omega}) &= |\det A| \cdot \frac{\hat{\phi}(A^T\omega)}{\hat{\phi}(\omega)} = |\det A| \cdot \frac{\hat{\rho}_{r,N}(A^T\omega)\zeta(A^T\omega)}{\hat{\rho}_{r,N}(\omega)\zeta(\omega)} \\ &= \frac{\zeta(A^T\omega)}{\zeta(\omega)} \cdot \frac{1}{(a^2+b^2)^{r-1}(a-ib)^N}, \quad \omega \in \mathbb{R}^2 \end{aligned}$$

Here, $\zeta(\omega) = (\mu(\omega))^{r+\frac{N}{2}}$ is the localizing multiplier for the isotropic polyharmonic B-spline in the case $N = 0$ and for the rotation-covariant polyharmonic B-splines in the case $N \in \mathbb{N}$; cf. (15)–(17).

The wavelet function ψ spanning a Riesz basis for the orthogonal complement

$$W_j = \overline{\text{span}\,\{2^{j/2}\psi(A^j \bullet -k), k \in \mathbb{Z}^2\}}^{L^2(\mathbb{R}^2)}$$

in $V_{j+1} = W_j \oplus V_j$ can also be gained in the frequency domain. For the quincunx dilation matrix A as in (18), a wavelet (or sometimes called a prewavelet, since the functions are not yet orthonormalized) is given by

$$\hat{\psi}(\omega) = G(e^{i\omega})\hat{\phi}(\omega) = e^{-i\omega_1}\overline{H(\omega + (\pi,\pi)^T)}M(\omega + (\pi,\pi)^T)\hat{\phi}(\omega),$$

compare with the section "Wavelets". The (pre)wavelet Riesz basis for W_j is then given by the family $\{\psi_{j,k} = 2^{j/2}\psi(A^j \bullet -k), k \in \mathbb{Z}^2\}$.

This basis in general is not orthonormal: $\langle\psi_{j,k}, \psi_{j,l}\rangle \neq \delta_{k,l}$. A function $f \in L^2(\mathbb{R}^2)$ can then be represented by the series

$$f = \sum\nolimits_{j\in\mathbb{Z},k\in\mathbb{Z}^2} \langle f, \tilde{\psi}_{j,k}\rangle \psi_{j,k} = \sum\nolimits_{j\in\mathbb{Z},k\in\mathbb{Z}^2} \langle f, \tilde{\psi}_{j,k}\rangle \psi_{j,k},$$

where $\{\tilde{\psi}_{j,k}\}_{k\in\mathbb{Z}^2}$ denotes the dual basis for each $j \in \mathbb{Z}$: $\langle \tilde{\psi}_{j,k}, \psi_{j,l}\rangle = \delta_{k,l}$. Its generator in the frequency domain is

$$\hat{\tilde{\psi}}(\omega) = e^{-i\omega_1}\overline{H(\omega + (\pi,\pi)^T)}\frac{M(\omega + (\pi,\pi)^T)}{M(A^T\omega)}\frac{\hat{\phi}(\omega)}{M(\omega)}.$$

In contrast, the formula

$$\hat{\Psi}(\omega) = \sqrt{\frac{M(\omega + (\pi,\pi)^T)}{M(A^T\omega)}}\hat{\psi}(\omega)$$

generates an orthonormal wavelet basis. It corresponds to the orthonormal basis of V_0 generated by the integer shifts of the orthonormalized scaling function Φ. These considerations show that there are three variants of a multiresolution implementation: an "orthonormal" one with respect to the orthonormalized scaling functions and corresponding orthonormal wavelets; one with the B-splines on the analysis side,

$$f = \sum\nolimits_{k\in\mathbb{Z}^2} \langle f, \phi_{j,k}\rangle \tilde{\phi}_{j,k} + \sum\nolimits_{k\in\mathbb{Z}^2} \langle f, \psi_{j,k}\rangle \tilde{\psi}_{j,k} \quad \text{for } f \in V_{j+1},$$

and finally one with the B-splines on the synthesis side:

$$f = \sum\nolimits_{k\in\mathbb{Z}^2} \langle f, \tilde{\phi}_{j,k}\rangle \phi_{j,k} + \sum\nolimits_{k\in\mathbb{Z}^2} \langle f, \tilde{\psi}_{j,k}\rangle \psi_{j,k} \quad \text{for } f \in V_{j+1}.$$

Both, the scaling filters $H(e^{i\omega})$ and the wavelet filters

$$G(e^{i\omega}) = e^{-i\omega_1}\overline{H(\omega + (\pi,\pi)^T)}M(\omega + (\pi,\pi)^T)$$

as well as their orthogonal and dual variants in our case are nonseparable and infinitely supported. Therefore, a spatial implementation of the decomposition would lead to truncation errors due to the necessary restriction to a finite number of samples. However, because of the closed form of H and therefore of G, the corresponding multiresolution decomposition or wavelet transform can be efficiently implemented in the frequency domain. The respective image first undergoes an FFT and then is filtered in the frequency domain by multiplication with the scaling filter H and the wavelet filter G. This method automatically imposes periodic boundary conditions.

The coefficients resulting from the high-pass filtering with G are the detail coefficients. They are stored, whereas the coefficients resulting from the low-pass filtering H are reconsidered for the next iteration step.

$$\sum\nolimits_{k\in\mathbb{Z}^2} \langle f, \phi_{j+1,k}\rangle \tilde{\phi}_{j+1,k} = \sum\nolimits_{k\in\mathbb{Z}^2} \langle f, \phi_{j,k}\rangle \tilde{\phi}_{j,k} + \sum\nolimits_{k\in\mathbb{Z}^2} \langle f, \psi_{j,k}\rangle \tilde{\psi}_{j,k}.$$

For the details and tricks of the frequency domain implementation, cf. [25, 48, 49, 71].

Figure 16 shows the multiresolution decomposition for the scaling function $\phi = \mathcal{R}_{\mu}^{2,1}$. There, it was assumed that the image is bandlimited and projected on the space V_0, which has the advantage that the coefficients do not depend on the chosen flavor of the scaling function, i.e., orthogonal, B-spline, or dual. Qualitatively, the transform is very similar to a multiscale gradient with the real part corresponding to the x_2 derivative and the imaginary part corresponding to the x_1 derivative [28].

6 Open Questions

In this chapter, a method for the construction of spline multiresolution bases was described. It yields a nice variety of new bases with several parameters for adaption and tuning. In the last decade, the notion of compressive sampling or compressed sensing arose, which is footing on the existence of well-adaptable bases. In fact, the idea behind compressed sensing is that certain functions have a sparse representation, if the underlying basis is smartly chosen. In this case, the function can be reconstructed from very few samples because of the prior knowledge of sparsity in this underlying basis. As a consequence, the knowledge on sparsity allows to sample such a signal at a rate significantly under the Nyquist rate. (The Shannon-Nyquist sampling theorem says that a signal must be sampled at least two times faster than the signal's bandwidth to avoid loss of information.)

In the last 40 years, and virtually explosively in the last 10 years, many important theoretical results were proven in this field, in particular by D. Donoho, E. Candès, J. Romberg, and T. Tao. For an introduction and references on compressed sensing, see, e.g., [2, 9] and the website [55].

Compressed sensing is based upon two fundamental concepts: that of incoherence and that of sparsity. Let $\{x_i\}_{i=1,\dots,N}$ be an orthonormal basis of the vector space V. Let $f = \sum_{i=1}^{N} s_i x_i$ with $s_i = \langle f, x_i\rangle$. The signal f is called k-sparse, if only k of the coefficients are nonzero, $k \in \mathbb{N}$.

A general linear measurement process for signals consists in computing $M < N$ inner products $y_j = \langle f, y_j\rangle$ for a collection of vectors $\{y_j\}_j$. In matrix form,

$$g = Yf = YXs,$$

where Y and X are the matrixes with $\{y_i\}_i$ and $\{x_j\}_j$ as columns, and YX is an $M \times N$ matrix. If the function families Y and X are incoherent, i.e., if the incoherence measure

$$\mu(Y,\ X) = \sqrt{N} \max_{1\le i,j\le N} |\langle y_i, x_j\rangle| \ \in \ \left[1, \sqrt{N}\right]$$

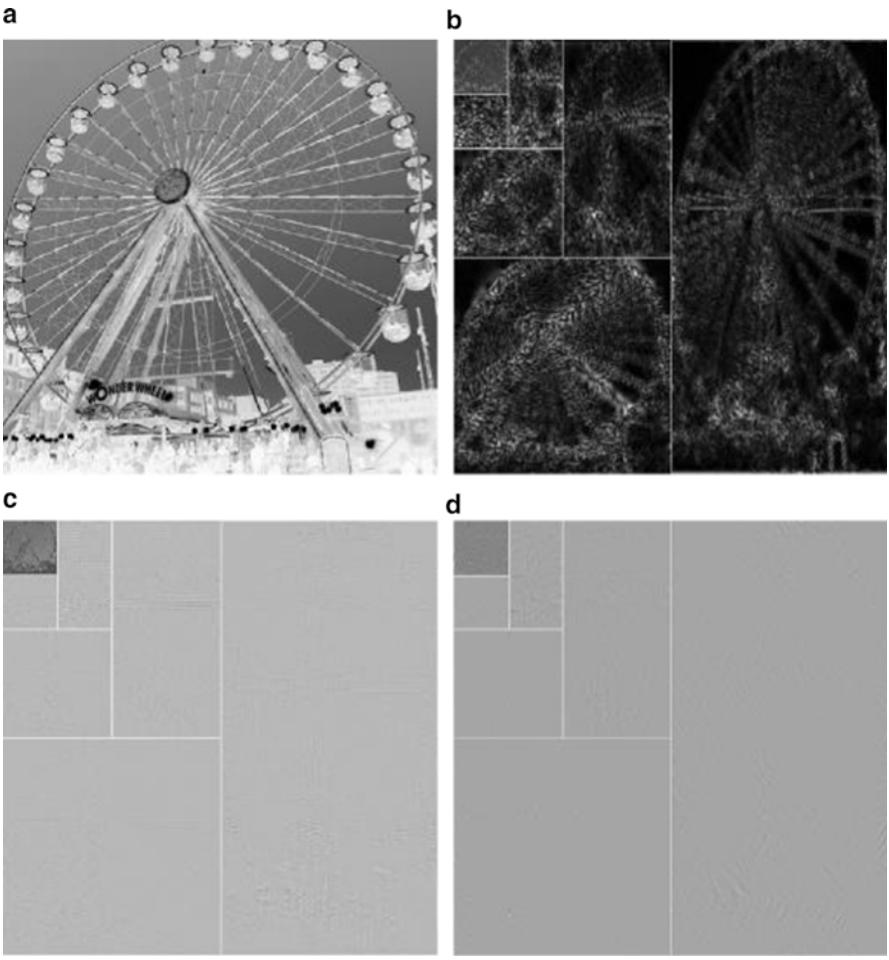


Fig. 16 Decomposition of an image [18, Part of IM115.tif] into approximation and wavelet coefficients. (**a**) Original image. (**b**) Matrix of the absolute values of the multiresolution coefficients. Large coefficients are *white*. The approximation coefficients are in the upper left band; the other bands are wavelet coefficients on six scales. (**c**) Real part of the coefficient matrix and (**d**) imaginary part of the decomposition matrix for $\phi = \mathcal{R}_\mu{}^{2,1}$ and the corresponding wavelets. The coefficients had their intensity rescaled for better contrast

is close to one, then under mild additional conditions, the k-sparse signal f can be reconstructed from $M > C\mu^2(Y, X)k \ln N$ samples with overwhelming probability.

Wavelet bases have proven to be very suitable for compressed sensing. It is an open question to classify the signals from certain applications and to estimate in which appropriate B-spline basis they have a k-sparse representation. Then adequate bases and function families incoherent with the spline bases have to be identified.

In the last years, the concept of sparsity entered image processing. It has proven to help immensely accelerate the solution of inverse problems and reconstruction algorithms, e.g., in medical imaging, such as in magnetic resonance imaging [42], computed tomography [12], photo-acoustic tomography [39], tomosynthesis [29], and others. In this area, as well as in other fields of imaging, it can be expected that the combination of splines – due to their easy modeling and the fast frequency domain algorithms – multiresolution and wavelets, and sparsity will lead to novel impressing fast algorithms for image reconstruction.

7 Conclusion

In the design procedure for scaling functions of multiresolution analyses, regularity and decay features, as well as symmetry properties, can be tuned by an appropriate modeling in frequency domain. The idea is to start in the frequency domain with a polynomial function P that fulfills the required symmetry features and that has a degree, such that $1/P$ decays sufficiently fast. This assures that the resulting scaling function has the desired regularity. However, $1/P$ in general is not an L^2 function and has to be multiplied with a localizing trigonometric polynomial ν that eliminates the zeros in the denominator such that $\frac{\nu}{P}$ becomes square integrable. The choice of this trigonometric polynomial has to be taken carefully to be compatible with the required features modeled in $1/P$. Then under mild additional conditions, the fraction

$$\hat{\phi} = \frac{\nu}{P}$$

is the scaling function of a multiresolution analysis. This construction can be performed for 1D and higher dimensional spaces likewise. In the time domain, the resulting scaling function is a piecewise polynomial, thus a spline. This design procedure for scaling functions unites the concepts of splines and of multiresolution.

Interestingly, the polynomial in the denominator can be of a fractional or a complex degree and therefore allows a fine tuning of the scaling function's properties. However, the scaling function then becomes an infinite series of shifted (truncated) polynomials. The numerical calculation with the approximating basis of the multiresolution analysis in the time domain would cause truncation errors, which is unfavorable. But due to the construction of ϕ in the frequency domain and due to the closed form there, the implementation in the frequency domain with periodic boundary conditions yields a fast and stable multiresolution algorithm suitable for image analysis tasks.

Cross-References

▸ Compressive Sensing
▸ Gabor Analysis for Imaging

▸ Neighborhood Filters and the Recovery of 3D Information
▸ Sampling Methods
▸ Starlet Transform in Astronomical Data Processing

References

1. Aldroubi, A., Unser, M.A. (eds.): Wavelets in Medicine and Biology. CRC, Boca Raton (1996)
2. Baraniuk, R.: Compressive Sensing. IEEE Signal Process. Mag. **4**(4), 118–120, 124 (2007)
3. Bartels, R.H., Bealty, J.C., Beatty, J.C.: An Introduction to Splines for Use in Computer Graphics and Geometric Modeling. Morgan Kaufman, Los Altos (1995)
4. Battle, G.: A block spin construction of ondelettes. Part 1: Lemarié functions. Commun. Math. Phys. **110**, 601–615 (1987)
5. Blu, T., Unser, M.: The fractional spline wavelet transform: definition and implementation. In: Proceedings of the 25th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'00), Istanbul, 5–9 June 2000, vol. 1, pp. 512–515 (2000)
6. Blu, T., Unser, M.: A complete family of scaling functions: the (α, τ)-fractional splines. In: Proceedings of the 28th International Conference on Acoustics, Speech, and Signal Processing (ICASSP'03), Hong Kong SAR, 6–10 Apr 2003, vol. 6, pp. 421–424 (2003)
7. Buhmann, M.D.: Radial Basis Functions: Theory and Implementations. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, Cambridge (2003)
8. Burt, P.J., Adelson, E.H.: The Laplacian pyramid as a compact image code. IEEE Trans. Commun. **31**(4), 532–540 (1983)
9. Candès, E.J., Wakin, M.B.: An introduction to compressive sampling. IEEE Signal Process. Mag. **25**(2), 21–30 (2008)
10. Champeney, D.C.: A Handbook of Fourier Theorems. Cambridge University Press, Cambridge (1987)
11. Chen, H.-l.: Complex Harmonic Splines, Periodic Quasi-wavelets, Theory and Applications. Kluwer Academic, Dordrecht (2000)
12. Choi, J.Y., Kim, M.W., Seong, W., Ye, J.C.: Compressed sensing metal artifact removal in dental CT. In: Proceedings of IEEE International Symposium on Biomedical Imaging (ISBI), 28 June–1 July 2009, Boston, pp. 334–337 (2009)
13. Christensen, O.: An Introduction to Frames and Riesz Bases. Birkhäuser, Boston (2003)
14. Christensen, O.: Frames and Bases: An Introductory Course. Applied and Numerical Harmonic Analysis. Birkhäuser, Boston (2008)
15. Chui, C.K.: Multivariate Splines. Society for Industrial and Applied Mathematics, Philadelphia (1988)
16. Chui, C.: Wavelets—a Tutorial in Theory and Practice. Academic, San Diego (1992)
17. Chui, C.K. (ed.): Wavelets: A Tutorial in Theory and Applications. Academic, Boston (1992)
18. Condat, L.: Image database. Online resource. http://www.greyc.ensicaen.fr/_lcondat/imagebase.html (2010). (Version of 22 Apr 2010)
19. Condat, L., Forster-Heinlein, B., Van De Ville, D.: A new family of rotation-covariant wavelets on the hexagonal lattice. In: SPIE Wavelets XII, San Diego, Aug 2007
20. Dahmen, W., Kurdila, A., Oswald, P. (eds.): Multiscale Wavelet Methods for Partial Differential Equations. Volume 6 of Wavelet Analysis and Its Applications. Academic, San Diego (1997)
21. Daubechies, I.: Ten Lectures on Wavelets. Society for Industrial and Applied Mathematics, Philadelphia (1992)
22. de Boor, C., Höllig, K., Riemenschneider, S.: Box Splines. Volume 98 of Applied Mathematical Sciences. Springer, New York (1993)

23. de Boor, C., Devore, R.A., Ron, A.: Approximation from shift invariant subspaces of $L_2(R^d)$. Trans. Am. Math. Soc. **341**(2), 787–806 (1994)
24. Dierckx, P.: Curve and Surface Fitting with Splines. McGraw-Hill, New York (1993)
25. Feilner, M., Van De Ville, D., Unser, M.: An orthogonal family of quincunx wavelets with continuously adjustable order. IEEE Trans. Image Process. **4**(4), 499–510 (2005)
26. Forster, B., Massopust, P.: Statistical encounters with complex B-splines. Constr. Approx. **29**(3), 325–344 (2009)
27. Forster, B., Blu, T., Unser, M.: Complex B-splines. Appl. Comput. Harmon. Anal. **20**(2), 261–282 (2006)
28. Forster, B., Blu, T., Van De Ville, D., Unser, M.: Shiftinvariant spaces from rotation-covariant functions. Appl. Comput. Harmon. Anal. **25**(2), 240–265 (2008)
29. Frikel, J.: A new framework for sparse regularization in limited angle x-ray tomography. In: IEEE International Symposium on Biomedical Imaging, Rotterdam (2010)
30. Giles, R.C., Kotiuga, P.R., Mansuripur, M.: Parallel micromagnetic simulations using Fourier methods on a regular hexagonal lattice. IEEE Trans. Magn. **7**(5), 3815–3818 (1991)
31. Grigoryan, A.M.: Efficient algorithms for computing the 2-D hexagonal Fourier transforms. IEEE Trans Signal Process. **50**(6), 1438–1448 (2002)
32. Hales, T.C.: The honeycomb conjecture. Discret. Comput. Geom. **25**, 1–22 (2001)
33. Heil, C., Walnut, D.F.: Fundamental Papers in Wavelet Theory, New edn. Princeton University Press, Princeton (2006)
34. Jones, D.S.: Generalised Functions. McGraw-Hill, London (1966)
35. Lai, M.-J., Schumaker, L.L.: Spline Functions on Triangulations. Cambridge University Press, Cambridge (2007)
36. Laine, A.F., Schuler, S., Fan, J., Huda, W.: Mammographic feature enhancement by multiscale analysis. IEEE Trans. Med. Imaging **13**(4), 725–740 (1994)
37. Legrand, P.: Local regularity and multifractal methods for image and signal analysis. In: Abry, P., Gonçalves, P., Véhel, L. (eds.) Scaling, Fractals and Wavelets, chap. 11. Wiley-ISTE, London (2009)
38. Lemarié, P.-G.: Ondelettes a localisation exponentielle. J. Math. Pures Appl. **67**, 227–236 (1988)
39. Lesage, F., Provost, J.: The application of compressed sensing for photo-acoustic tomography. IEEE Trans. Med. Imaging **28**(4), 585–594 (2009)
40. Lipow, P.R., Schoenberg, I.J.: Cardinal interpolation and spline functions. III: Cardinal hermite interpolation. Linear Algebra Appl. **6**, 273–304 (1973)
41. Louis, A.K., Maaß, P., Rieder, A.: Wavelets: Theory and Applications. Wiley, New York (1997)
42. Lustig, M., Donoho, D., Pauly, J.M.: Sparse MRI: the application of compressed sensing for rapid MR imaging. Magn. Reson. Med. **58**(6), 1182–1195 (2007)
43. Mallat, S.: Multiresolution approximations and wavelet orthonormal bases of $L^2(R)$. Trans. Am. Math. Soc. **315**, 69–87 (1989)
44. Mallat, S.G.: A Wavelet Tour of Signal Processing. Academic, San Diego (1998)
45. Mersereau, R.M.: The processing of hexagonally sampled two-dimensional signals. Proc. IEEE **67**(6), 930–949 (1979)
46. Meyer, Y.: Wavelets and Operators. Cambridge University Press, Cambridge (1992)
47. Middleton, L., Sivaswamy, J.: Hexagonal Image Processing: A Practical Approach. Advances in Pattern Recognition. Springer, Berlin (2005)
48. Nicolier, F., Laligant, O., Truchetet, F.: B-spline quincunx wavelet transforms and implementation in Fourier domain. Proc. SPIE **3522**, 223–234 (1998)
49. Nicolier, F., Laligant, O., Truchetet, F.: Discrete wavelet transform implementation in Fourier domain for multidimensional signal. J. Electron. Imaging **11**, 338–346 (2002)
50. Nürnberger, G.: Approximation by Spline Functions. Springer, Berlin (1989)
51. Plonka, G., Tasche, M.: On the computation of periodic spline wavelets. Appl. Comput. Harmon. Anal. **2**, 1–14 (1995)
52. Püschel, M., Rötteler, M.: Algebraic signal processing theory: 2D spatial hexagonal lattice. IEEE Trans. Image Proc. **16**(6), 1506–1521 (2007)

53. Rabut, C.: Elementary m-harmonic cardinal B-splines. Numer. Algorithms **2**, 39–62 (1992)
54. Rabut, C.: High level m-harmonic cardinal B-splines. Numer. Algorithms **2**, 63–84 (1992)
55. Rice University: Compressive sensing resources. Online resource. http://dsp.rice.edu/cs (2010). (Version of 30 Apr 2010)
56. Rudin, W.: Functional Analysis. International Series in Pure and Applied Mathematics. McGraw-Hill, New York (1991)
57. Sablonnière, P., Sbibih, D.: B-splines with hexagonal support on a uniform three-direction mesh of the plane. C. R. Acad. Sci. Paris Sér. I **319**, 227–282 (1994)
58. Schempp, W.: Complex Contour Integral Representation of Cardinal Spline Functions. Volume 7 of Contemporary Mathematics. American Mathematical Society, Providence (1982)
59. Schoenberg, I.J.: Contributions to the problem of approximation of equidistant data by analytic functions. Part A – on the problem of osculatory interpolation. A second class of analytic approximation formulae. Q. Appl. Math. **4**, 112–141 (1946)
60. Schoenberg, I.J.: Contributions to the problem of approximation of equidistant data by analytic functions. Part A – on the problem of smoothing or graduation. A first class of analytic approximation formulae. Q. Appl. Math. **4**, 45–99 (1946)
61. Schoenberg, I.J.: Cardinal interpolation and spline functions. J. Approx. Theory **2**, 167–206 (1969)
62. Schoenberg, I.J.: Cardinal interpolation and spline functions. II: interpolation of data of power growth. J. Approx. Theory **6**, 404–420 (1972)
63. Schwartz, L.: Théorie des distributions. Hermann, Paris (1998)
64. Unser, M.: Splines: a perfect fit for medical imaging. In: Sonka, M., Fitzpatrick, J.M. (eds.) Progress in Biomedical Optics and Imaging, vol. 3, no. 22, vol. 4684, Part I of Proceedings of the SPIE International Symposium on Medical Imaging: Image Processing (MI'02), San Diego, 24–28 Feb 2002, pp. 225–236
65. Unser, M., Blu, T.: Fractional splines and wavelets. SIAM Rev. **42**(1), 43–67 (2000)
66. Unser, M., Aldroubi, A., Eden, M.: Fast B-spline transforms for continuous image representation and interpolation. IEEE Trans. Pattern Anal. Mach. Intell. **13**(3), 277–285 (1991)
67. Unser, M., Aldroubi, A., Eden, M.: On the asymptotic convergence of B-spline wavelets to Gabor functions. IEEE Trans. Inf. Theory **38**, 864–872 (1992)
68. Unser, M., Aldroubi, A., Eden, M.: B-spline signal processing: part I—Theory. IEEE Trans. Signal Process. **41**(2), 821–833 (1993)
69. Unser, M., Aldroubi, A., Eden, M.: B-spline signal processing: part II—Efficient design and applications. IEEE Trans. Signal Process. **41**(2), 834–848 (1993)
70. Van De Ville, D., Blu, T., Unser, M., Philips, W., Lemahieu, I., Van de Walle, R.: Hex-splines: a novel spline family for hexagonal lattices. IEEE Trans. Image Process. **13**(6), 758–772 (2004)
71. Van De Ville, D., Blu, T., Unser, M.: Isotropic polyharmonic B-splines: scaling functions and wavelets. IEEE Trans. Image Process. **14**(11), 1798–1813 (2005)
72. Watson, A.B., Ahumuda, A.J., Jr.: Hexagonal orthogonal-oriented pyramid as a model of image representation in visual cortex. IEEE Trans. Biomed. Eng. **36**(1), 97–106 (1989)
73. Wendt, H., Roux, S.G., Jaffard, S., Abry, P.: Wavelet leaders and bootstrap for multifractal analysis of images. Signal Process. **89**, 1100–1114 (2009)
74. Wojtaszczyk, P.: A Mathematical Introduction to Wavelets. Volume 37 of London Mathematical Society Student Texts. Cambridge University Press, Cambridge (1997)
75. Young, R.: An Introduction to Nonharmonic Fourier Series. Academic, New York (1980) (revised first edition 2001)

Gabor Analysis for Imaging

Ole Christensen, Hans G. Feichtinger, and Stephan Paukner

Contents

1 Introduction 1718
2 Tools from Functional Analysis 1718
The Pseudo-inverse Operator 1719
Bessel Sequences in Hilbert Spaces 1720
General Bases and Orthonormal Bases 1721
Frames and Their Properties 1722
3 Operators 1724
The Fourier Transform 1724
Translation and Modulation 1726
Convolution, Involution, and Reflection 1727
The Short-Time Fourier Transform 1727
4 Gabor Frames in $\mathbf{L^2(\mathbb{R}d)}$ 1730
5 Discrete Gabor Systems 1734
Gabor Frames in $\ell^2(\mathbb{Z})$ 1734
Finite Discrete Periodic Signals 1735
Frames and Gabor Frames in $\mathbb{C}^L$ 1736
6 Image Representation by Gabor Expansion 1739
2D Gabor Expansions 1739
Separable Atoms on Fully Separable Lattices 1742
Efficient Gabor Expansion by Sampled STFT 1746
Visualizing a Sampled STFT of an Image 1747

O. Christensen (✉)
Department of Mathematics, Technical University of Denmark, Lyngby, Denmark
e-mail: Ole.Christensen@mat.dtu.dk

H.G. Feichtinger
University of Vienna, Vienna, Austria
e-mail: hans.feichtinger@univie.ac.at

S. Paukner
Applied Research Center Communication Systems, GmbH, Vienna, Austria
e-mail: stephan+math@paukner.cc

© Springer Science+Business Media New York 2015

O. Scherzer (ed.), *Handbook of Mathematical Methods in Imaging*,
DOI 10.1007/978-1-4939-0790-8_54

Non-separable Atoms on Fully Separable Lattices 1751
7 Historical Notes and Hint to the Literature 1752
Cross-References 1753
References 1753

1 Introduction

In contrast to classical Fourier analysis, time–frequency analysis is concerned with *localized Fourier transforms*. Gabor analysis is an important branch of time–frequency analysis. Although significantly different, it shares with the wavelet transform methods the ability to describe the smoothness of a given function in a location-dependent way.

The main tool is the *sliding window Fourier transform* or *short-time Fourier transform* (STFT) in the context of audio signals. It describes the correlation of a signal with the time–frequency shifted copies of a fixed function (or window or atom). Thus, it characterizes a function by its transform over phase space, which is the time–frequency plane (TF-plane) in a musical context or the location–wave-number domain in the context of image processing.

Since the transition from the signal domain to the phase space domain introduces an enormous amount of data redundancy, suitable subsampling of the continuous transform allows for complete recovery of the signal from the sampled STFT. The knowledge about appropriate choices of windows and sampling lattices has increased significantly during the last three decades. Since the suggestion goes back to the idea of D. Gabor [45], this branch of TF analysis is called *Gabor analysis*. Gabor expansions are not only of interest due to their very natural interpretation but also algorithmically convenient due to a good understanding of algebraic and analytic properties of Gabor families.

In this chapter, we describe some of the generalities relevant for an understanding of Gabor analysis of functions on $\mathbb{R}^d$. We pay special attention to the case $d = 2$, which is the most important case for image processing and image analysis applications.

The chapter is organized as follows. Section 2 presents central tools from functional analysis in Hilbert spaces, e.g., the pseudo-inverse of a bounded operator and the central facts from frame theory. In Sect. 3, we introduce several operators that play important roles in Gabor analysis. Gabor frames on $L^2(\mathbb{R}^d)$ are introduced in Sect. 4, and their discrete counterpart are treated in Sect. 5. Finally, the application of Gabor expansions to image representation is considered in Sect. 6.

2 Tools from Functional Analysis

In this section, we recall basic facts from functional analysis. Unless another reference is given, a proof can be found in [17]. In the entire section, $\mathcal{H}$ denotes a separable Hilbert space with inner product $\langle \cdot, \cdot \rangle$.

The Pseudo-inverse Operator

It is well known that an arbitrary matrix has a pseudo-inverse, which can be used to find the minimal-norm least squares solution of a linear system. In the case of an operator on infinite-dimensional Hilbert spaces, one has to restrict the attention to linear operators with *closed range* in order to obtain a pseudo-inverse. Observe that a bounded operator (We will always assume *linearity*!) U on a Hilbert space $\mathcal{H}$ is invertible if and only if it is injective and surjective, while injectivity combined with a dense range is not sufficient in the infinite-dimensional case. However, *if* the range of U is closed, there exists a "*right-inverse operator*" $U^\dagger$ in the following sense:

Lemma 1. *Let $\mathcal{H}, \mathcal{K}$ be Hilbert spaces, and suppose that $U : \mathcal{K} \to \mathcal{H}$ is a bounded operator with closed range $\mathcal{R}_U$. Then there exists a bounded operator $U^\dagger : \mathcal{H} \to \mathcal{K}$ for which*

$$UU^\dagger x = x, \ \forall x \in \mathcal{R}_U. \tag{1}$$

Proof. Consider the operator obtained by taking the restriction of U to the orthogonal complement of the kernel of U, i.e., let

$$\tilde{U} := U_{|\mathcal{N}_U^\perp} : \mathcal{N}_U^\perp \to \mathcal{H}.$$

Obviously, $\tilde{U}$ is linear and bounded. $\tilde{U}$ is also injective: if $\tilde{U}x = 0$, it follows that $x \in \mathcal{N}_U^\perp \cap \mathcal{N}_U = \{0\}$. We prove next that the range of $\tilde{U}$ equals the range of U. Given $y \in \mathcal{R}_U$, there exists $x \in \mathcal{K}$ such that $Ux = y$. By writing $x = x_1 + x_2$, where $x_1 \in \mathcal{N}_U^\perp, \ x_2 \in \mathcal{N}_U$, we obtain that

$$\tilde{U}x_1 = Ux_1 = U(x_1 + x_2) = Ux = y.$$

It follows from Banach's theorem that $\tilde{U}$ has a bounded inverse

$$\tilde{U}^{-1} : \mathcal{R}_U \to \mathcal{N}_U^\perp.$$

Extending $\tilde{U}^{-1}$ by zero on the orthogonal complement of $\mathcal{R}_U$, we obtain a bounded operator $U^\dagger : \mathcal{H} \to \mathcal{K}$ for which $UU^\dagger x = x$ for all $x \in \mathcal{R}_U$. ■

The operator $U^\dagger$ constructed in the proof of Lemma 1 is called the *pseudo-inverse* of U. In the literature, one will often see the pseudo-inverse of an operator U defined as the unique operator $U^\dagger$ satisfying that

$$\mathcal{N}_{U^\dagger} = \mathcal{R}_U^\perp, \ \ \mathcal{R}_{U^\dagger} = \mathcal{N}_U^\perp, \text{ and } UU^\dagger x = x, x \in \mathcal{R}_U; \tag{2}$$

this definition is equivalent to the above construction. We collect some properties of $U^\dagger$ and its relationship to U.

Lemma 2. *Let $U : \mathcal{K} \to \mathcal{H}$ be a bounded operator with closed range. Then the following holds:*

(i) *The orthogonal projection of $\mathcal{H}$ onto $\mathcal{R}_U$ is given by $UU^\dagger$.*
(ii) *The orthogonal projection of $\mathcal{K}$ onto $\mathcal{R}_{U^\dagger}$ is given by $U^\dagger U$.*
(iii) *U^* has closed range, and $(U^*)^\dagger = (U^\dagger)^*$.*
(iv) *On $\mathcal{R}_U$, the operator $U^\dagger$ is given explicitly by*

$$U^\dagger = U^*(UU^*)^{-1}. \tag{3}$$

Bessel Sequences in Hilbert Spaces

When we deal with infinite-dimensional vector spaces, we need to consider expansions in terms of infinite series. The purpose of this section is to introduce a condition that ensures that the relevant infinite series actually converge. When speaking about a *sequence* $\{f_k\}_{k=1}^\infty$ in $\mathcal{H}$, we mean an *ordered* set, i.e., $\{f_k\}_{k=1}^\infty = \{f_1, f_2, \dots\}$. That we have chosen to index the sequence by the natural numbers is just for convenience.

Definition 1. A sequence $\{f_k\}_{k=1}^\infty$ in $\mathcal{H}$ is called a Bessel sequence if there exists a constant $B > 0$ such that

$$\sum_{k=1}^{\infty} |\langle f, f_k\rangle|^2 \le B\, ||f||^2, \ \forall f \in \mathcal{H}. \tag{4}$$

Any number B satisfying (4) is called a *Bessel bound* for $\{f_k\}_{k=1}^\infty$. The *optimal bound* for a given Bessel sequence $\{f_k\}_{k=1}^\infty$ is the smallest possible value of $B > 0$ satisfying (4). Except for the case $f_k = 0, \ \forall k \in \mathbb{N}$, the optimal bound always exists.

Theorem 1. *Let $\{f_k\}_{k=1}^\infty$ be a sequence in $\mathcal{H}$ and $B > 0$ be given. Then $\{f_k\}_{k=1}^\infty$ is a Bessel sequence with Bessel bound B if and only if*

$$T : \{c_k\}_{k=1}^\infty \to \sum_{k=1}^{\infty} c_k f_k$$

defines a bounded operator from $\ell^2(\mathbb{N})$ into $\mathcal{H}$ and $||T|| \le \sqrt{B}$.

The operator T is called the *synthesis operator*. The adjoint T^* is called the *analysis operator* and is given by

$$T^* : \mathcal{H} \to \ell^2(\mathbb{N}), \ \ T^* f = \{\langle f, f_k\rangle\}_{k=1}^\infty.$$

These operators play key roles in the theory of frames, to be considered in section "Frames and Their Properties."

The Bessel condition (4) remains the same, regardless of how the elements $\{f_k\}_{k=1}^{\infty}$ are numbered. This leads to a very important consequence of Theorem 1:

Corollary 1. *If $\{f_k\}_{k=1}^{\infty}$ is a Bessel sequence in $\mathcal{H}$, then $\sum_{k=1}^{\infty} c_k f_k$ converges unconditionally for all $\{c_k\}_{k=1}^{\infty} \in \ell^2(\mathbb{N})$, i.e., the series is convergent, irrespective of how and in which order the summation is realized.*

Thus, a reordering of the elements in $\{f_k\}_{k=1}^{\infty}$ will not affect the series $\sum_{k=1}^{\infty} c_k f_k$ when $\{c_k\}_{k=1}^{\infty}$ is reordered the same way: the series will converge toward the same element as before. For this reason, we can choose an arbitrary indexing of the elements in the Bessel sequence; in particular, it is not a restriction that we present all results with the natural numbers as index set. As we will see in the sequel, all orthonormal bases and frames are Bessel sequences.

General Bases and Orthonormal Bases

We will now briefly consider bases in Hilbert spaces. In particular, we will discuss orthonormal bases, which are the infinite-dimensional counterparts of the canonical bases in $\mathbb{C}^n$. Orthonormal bases are widely used in mathematics as well as physics, signal processing, and many other areas where one needs to represent functions in terms of "elementary building blocks."

Definition 2. Consider a sequence $\{e_k\}_{k=1}^{\infty}$ of vectors in $\mathcal{H}$.

(i) The sequence $\{e_k\}_{k=1}^{\infty}$ is a (Schauder) basis for $\mathcal{H}$ if for each $f \in \mathcal{H}$, there exist unique scalar coefficients $\{c_k(f)\}_{k=1}^{\infty}$ such that

$$f = \sum_{k=1}^{\infty} c_k(f) e_k. \tag{5}$$

(ii) A basis $\{e_k\}_{k=1}^{\infty}$ is an unconditional basis if the series (5) converges unconditionally for each $f \in \mathcal{H}$.

(iii) A basis $\{e_k\}_{k=1}^{\infty}$ is an orthonormal basis if $\{e_k\}_{k=1}^{\infty}$ is an orthonormal system, i.e., if

$$\langle e_k, e_j \rangle = \delta_{k,j} = \begin{cases} 1 & \text{if} \quad k = j, \\ 0 & \text{if} \quad k \neq j. \end{cases}$$

An orthonormal basis leads to an expansion of the type (5) with an explicit expression for the coefficients $c_k(f)$:

Theorem 2. *If $\{e_k\}_{k=1}^{\infty}$ is an orthonormal basis, then each $f \in \mathcal{H}$ has an unconditionally convergent expansion*

$$f = \sum_{k=1}^{\infty} \langle f, e_k \rangle e_k. \tag{6}$$

In practice, orthonormal bases are certainly the most convenient bases to use: for other types of bases, the representation (6) has to be replaced by a more complicated expression. Unfortunately, the conditions for $\{e_k\}_{k=1}^{\infty}$ being an orthonormal basis are strong, and often it is impossible to construct orthonormal bases satisfying extra conditions. We discuss this in more detail later. Note also that it is not always a good idea to use the Gram–Schmidt orthonormalization procedure to construct an orthonormal basis from a given basis: it might destroy special properties of the basis at hand. For example, the special structure of a Gabor basis (to be discussed later) will be lost.

Frames and Their Properties

We are now ready to introduce one of the central subjects:

Definition 3. A sequence $\{f_k\}_{k=1}^{\infty}$ of elements in $\mathcal{H}$ is a frame for $\mathcal{H}$ if there exist constants $A, B > 0$ such that

$$A \, ||f||^2 \leq \sum_{k=1}^{\infty} |\langle f, f_k \rangle|^2 \leq B \, ||f||^2, \quad \forall f \in \mathcal{H}. \tag{7}$$

The numbers A and B are called *frame bounds*. A special role is played by frames for which the optimal frame bounds coincide:

Definition 4. A sequence $\{f_k\}_{k=1}^{\infty}$ in $\mathcal{H}$ is a tight frame if there exists a number $A > 0$ such that

$$\sum_{k=1}^{\infty} |\langle f, f_k \rangle|^2 = A \, ||f||^2, \quad \forall f \in \mathcal{H}.$$

The number A is called the frame bound.

Since a frame $\{f_k\}_{k=1}^{\infty}$ is a Bessel sequence, the operator

$$T : \ell^2(\mathbb{N}) \to \mathcal{H}, \; T\{c_k\}_{k=1}^{\infty} = \sum_{k=1}^{\infty} c_k f_k \tag{8}$$

is bounded by Theorem 1. Composing T and T^*, we obtain the *frame operator*

$$S : \mathcal{H} \to \mathcal{H}, \quad Sf = TT^* f = \sum_{k=1}^{\infty} \langle f, f_k \rangle f_k. \tag{9}$$

The *frame decomposition*, stated in (10) below, is the most important frame result. It shows that if $\{f_k\}_{k=1}^{\infty}$ is a frame for $\mathcal{H}$, then every element in $\mathcal{H}$ has a representation as an infinite linear combination of the frame elements. Thus, it is natural to view a frame as a "generalized basis."

Theorem 3. *Let $\{f_k\}_{k=1}^{\infty}$ be a frame with frame operator S. Then*

$$f = \sum_{k=1}^{\infty} \langle f, S^{-1} f_k \rangle f_k, \quad \forall f \in \mathcal{H}, \tag{10}$$

and

$$f = \sum_{k=1}^{\infty} \langle f, f_k \rangle S^{-1} f_k, \quad \forall f \in \mathcal{H}. \tag{11}$$

Both series converge unconditionally for all $f \in \mathcal{H}$.

Theorem 3 shows that all information about a given vector $f \in \mathcal{H}$ is contained in the sequence $\{\langle f, S^{-1} f_k \rangle\}_{k=1}^{\infty}$. The numbers $\langle f, S^{-1} f_k \rangle$ are called *frame coefficients*. The sequence $\{S^{-1} f_k\}_{k=1}^{\infty}$ is also a frame; it is called the *canonical dual frame* of $\{f_k\}_{k=1}^{\infty}$.

Theorem 3 also immediately reveals one of the main difficulties in frame theory. In fact, in order for the expansions (10) and (11) to be applicable in practice, we need to be able to find the operator S^{-1} or at least to calculate its action on all f_k, $k \in \mathbb{N}$. In general, this is a major problem. One way of circumventing the problem is to consider only tight frames:

Corollary 2. *If $\{f_k\}_{k=1}^{\infty}$ is a tight frame with frame bound A, then the canonical dual frame is $\{A^{-1} f_k\}_{k=1}^{\infty}$, and*

$$f = \frac{1}{A} \sum_{k=1}^{\infty} \langle f, f_k \rangle f_k, \quad \forall f \in \mathcal{H}. \tag{12}$$

By a suitable scaling of the vectors $\{f_k\}_{k=1}^{\infty}$ in a tight frame, we can always obtain that $A = 1$; in that case, (12) has exactly the same form as the representation via an orthonormal basis; see (6). Thus, such frames can be used without any additional computational effort compared with the use of orthonormal bases; however, the family does not have to be linearly independent now.

Tight frames have other advantages. For the design of frames with prescribed properties, it is essential to control the behavior of the canonical dual frame, but the complicated structure of the frame operator and its inverse makes this difficult. If, e.g., we consider a frame $\{f_k\}_{k=1}^{\infty}$ for $L^2(\mathbb{R})$ consisting of functions with exponential decay, nothing guarantees that the functions in the canonical dual frame $\{S^{-1}f_k\}_{k=1}^{\infty}$ have exponential decay. However, for tight frames, questions of this type trivially have satisfactory answers, because the dual frame equals the original one. Also, for a tight frame, the canonical dual frame automatically has the same structure as the frame itself: if the frame has Gabor structure (to be described in Sect. 4), the same is the case for the canonical dual frame.

There is another way to avoid the problem of inverting the frame operator S. A frame that is *not* a basis is said to be *overcomplete;* in the literature, the term *redundant frame* is also used. For frames $\{f_k\}_{k=1}^{\infty}$ that are *not* bases, one can replace the canonical dual $\{S^{-1}f_k\}_{k=1}^{\infty}$ by other frames:

Theorem 4. *Assume that $\{f_k\}_{k=1}^{\infty}$ is an overcomplete frame. Then there exist frames $\{g_k\}_{k=1}^{\infty} \neq \{S^{-1}f_k\}_{k=1}^{\infty}$ for which*

$$f = \sum_{k=1}^{\infty} \langle f, g_k \rangle f_k, \quad \forall f \in \mathcal{H}. \tag{13}$$

A frame $\{g_k\}_{k=1}^{\infty}$ satisfying (13) is called a *dual frame* of $\{f_k\}_{k=1}^{\infty}$. The hope is to find dual frames that are easier to calculate or have better properties than the canonical dual. Examples of this type can be found in [17].

3 Operators

In this section, we introduce several operators that play key roles in Gabor analysis. In particular, we will need the basic properties of the *localized Fourier transform*, which is called the STFT (short-time Fourier transform). It is natural for us to start with the *Fourier transform*, which is defined as an integral transform on the space of all (Lebesgue) integrable functions, denoted by $\mathrm{L}^1(\mathbb{R}^d)$.

The Fourier Transform

Definition 5. For $f \in \mathrm{L}^1(\mathbb{R}^d)$, the *Fourier transform* is defined as

$$\hat{f}(\omega) := (\mathcal{F}f)(\omega) := \int_{\mathbb{R}^d} f(x)\, e^{-2\pi \mathrm{i} x \cdot \omega}\, dx, \tag{14}$$

where $x \cdot \omega = \sum_{k=1}^{d} x_k \omega_k$ is the usual scalar product of vectors in $\mathbb{R}^d$.

Lemma 3 (Riemann–Lebesgue). *If $f \in \mathrm{L}^1(\mathbb{R}^d)$, then $\hat{f}$ is uniformly continuous and* $\lim_{|\omega|\to\infty} |\hat{f}(\omega)| = 0$.

The Fourier transform yields a continuous bijection from the Schwartz space $\mathcal{S}(\mathbb{R}^d)$ to $\mathcal{S}(\mathbb{R}^d)$. This follows from the fact that it turns analytic operations (differentiation) into multiplication with polynomials and vice versa:

$$\mathcal{F}(D^\alpha f) = (2\pi \mathrm{i})^{|\alpha|} X^\alpha (\mathcal{F} f) \tag{15}$$

and

$$D^\alpha (\mathcal{F} f) = (-2\pi \mathrm{i})^{|\alpha|} \mathcal{F}(X^\alpha f), \tag{16}$$

with a multi-index $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}_0^d$, $|\alpha| := \sum_{i=1}^d \alpha_i$, D^α as differential operator

$$D^\alpha f(x) := \frac{\partial^{\alpha_1} \cdots \partial^{\alpha_d}}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}} f(x_1, \ldots, x_d)$$

and X^α as multiplication operator $(X^\alpha f)(x) := x_1^{\alpha_1} \cdots x_d^{\alpha_d} f(x_1, \ldots, x_d)$. It follows from the definition that $\mathcal{S}(\mathbb{R}^d)$ is invariant under these operations, i.e.,

$$X^\alpha f \in \mathcal{S}(\mathbb{R}^d) \quad \text{and} \quad D^\alpha f \in \mathcal{S}(\mathbb{R}^d) \quad \forall \alpha \in \mathbb{N}_0^d \quad \forall f \in \mathcal{S}(\mathbb{R}^d).$$

Using the reflection operator $(\mathcal{I} f)(x) := f(-x)$, one can show that $\mathcal{F}^2 = \mathcal{I}$ and so $\mathcal{F}^4 = \mathrm{Id}_{\mathcal{S}(\mathbb{R}^d)}$. This yields

$$\mathcal{F}^{-1} = \mathcal{I}\mathcal{F} \tag{17}$$

and we can give an inversion formula explicitly:

Theorem 5 (Inversion Formula). *The Fourier transform is a bijection from $\mathcal{S}(\mathbb{R}^d)$ to $\mathcal{S}(\mathbb{R}^d)$, and the inverse operator is given by*

$$(\mathcal{F}^{-1} f)(x) = \int_{\mathbb{R}^d} f(\omega)\, e^{2\pi \mathrm{i} x \cdot \omega}\, d\omega \quad \forall x \in \mathbb{R}^d. \tag{18}$$

Furthermore,

$$\langle \mathcal{F} f, \mathcal{F} g \rangle_{\mathrm{L}^2} = \langle f, g \rangle_{\mathrm{L}^2} \quad \forall f, g \in \mathcal{S}(\mathbb{R}^d).$$

We can extend the Fourier transform to an isometric operator on all of $\mathrm{L}^2(\mathbb{R}^d)$. We will use the same symbol $\mathcal{F}$ although the Fourier transform on $\mathrm{L}^2(\mathbb{R}^d)$ is not defined by a Lebesgue integral (14) anymore if $f \in \mathrm{L}^2 \backslash \mathrm{L}^1(\mathbb{R}^d)$, but rather by

means of summability methods. Moreover, $\mathcal{F}f$ should be viewed as an equivalence class of functions, rather than a pointwise given function.

Theorem 6 (Plancherel). *If* $f \in \mathrm{L}^1 \cap \mathrm{L}^2(\mathbb{R}^d)$, *then*

$$\|f\|_{\mathrm{L}^2} = \|\mathcal{F}f\|_{\mathrm{L}^2}. \tag{19}$$

As a consequence, $\mathcal{F}$ *extends in a unique way to a unitary operator on* $\mathrm{L}^2(\mathbb{R}^d)$ *that satisfies Parseval's formula*

$$\langle f, g\rangle_{\mathrm{L}^2} = \langle \mathcal{F}f, \mathcal{F}g\rangle_{\mathrm{L}^2} \quad \forall f, g \in \mathrm{L}^2(\mathbb{R}^d). \tag{20}$$

In signal analysis, the isometry of the Fourier transform has the interpretation that it preserves the energy of a signal. For more details on the role of the Schwartz class for the Fourier transform, see [78, V].

Translation and Modulation

Definition 6. For $x, \omega \in \mathbb{R}^d$, we define the *translation operator* T_x by

$$(T_x f)(t) := f(t - x) \tag{21}$$

and the *modulation operator* M_ω by

$$(M_\omega f)(t) := e^{2\pi \mathrm{i}\omega \cdot t} f(t). \tag{22}$$

One has $T_x^{-1} = T_{-x}$ and $M_\omega^{-1} = M_{-\omega}$. The operator T_x is called a *time shift* and M_ω a *frequency shift.* Operators of the form $T_x M_\omega$ or $M_\omega T_x$ are called *time–frequency shifts* (TF-shifts). They satisfy the *commutation relations*

$$T_x M_\omega = e^{-2\pi \mathrm{i}x \cdot \omega} M_\omega T_x. \tag{23}$$

Time–frequency shifts are isometries on L^p for all $1 \le p \le \infty$, i.e.,

$$\|T_x M_\omega f\|_{\mathrm{L}^p} = \|f\|_{\mathrm{L}^p}.$$

The interplay of TF-shifts with the Fourier transform is as follows:

$$\widehat{T_x f} = M_{-x}\hat{f} \quad \text{or} \quad \mathcal{F}T_x = M_{-x}\mathcal{F} \tag{24}$$

and

$$\widehat{M_\omega f} = T_\omega \hat{f} \quad \text{or} \quad \mathcal{F}M_\omega = T_\omega \mathcal{F}. \tag{25}$$

Equation (25) explains why modulations are also called *frequency shifts*: modulations become translations on the Fourier transform side. Altogether, we have

$$\widehat{T_x M_\omega f} = M_{-x} T_\omega \hat{f} = e^{-2\pi i x \cdot \omega} T_\omega M_{-x} \hat{f}.$$

Convolution, Involution, and Reflection

Definition 7. The *convolution* of two functions $f, g \in \mathrm{L}^1(\mathbb{R}^d)$ is the function $f * g$ defined by

$$(f * g)(x) := \int_{\mathbb{R}^d} f(y)\, g(x-y)\, dy. \tag{26}$$

It satisfies

$$\|f * g\|_{\mathrm{L}^1} \le \|f\|_{\mathrm{L}^1} \|g\|_{\mathrm{L}^1} \quad \text{and} \quad \widehat{f * g} = \hat{f} \cdot \hat{g}.$$

One may view $f * g$ as f being "smeared" by g and vice versa. One can thus smoothen a function by convolving it with a narrow bump function.

Definition 8. The *involution* of a function is defined by

$$f^*(x) := \overline{f(-x)}. \tag{27}$$

It follows that

$$\widehat{f^*} = \bar{\hat{f}} \quad \text{and} \quad \widehat{\mathcal{I} f} = \mathcal{I} \hat{f}.$$

Finally, let us mention that convolution corresponds to pointwise multiplication (and conversely), i.e., the so-called *convolution theorem* is valid:

$$\widehat{g * f} = \hat{f} \cdot \hat{g}. \tag{28}$$

The Short-Time Fourier Transform

The Fourier transform as described in section "The Fourier Transform" provides only global frequency information of a signal f. This is useful for signals that do not vary during the time, e.g., for analyzing the spectrum of a violin tone. However, dynamic signals such as a melody have to be split into short time intervals over which it can be well approximated by a linear combination of few pure frequencies. Since sharp cutoffs would introduce discontinuities in the localized signal and

therefore leaking of the frequency spectrum, a smooth window function g is usually used in the definition of the short-time Fourier transform.

In image processing, one has plane waves instead of pure frequencies; thus, the global Fourier transform is only well suited to stripe-like patterns. Again, a localized version of the Fourier transform allows to determine dominant plane waves locally, and one can reconstruct an image from such a redundant transform. Gabor analysis deals with the question of how one can reconstruct an image from only somewhat overlapping local pieces, which are stored only in the form of a sampled (local) 2D Fourier transform (Fig. 1).

Definition 9. Fix a *window function* $g \in \mathrm{L}^2(\mathbb{R}^d)\backslash\{0\}$. The *short-time Fourier transform* (STFT), also called *(continuous) Gabor transform* of a function $f \in \mathrm{L}^2(\mathbb{R}^d)$ with respect to g, is defined as

$$(\mathcal{V}_g f)(x,\omega) := \int_{\mathbb{R}^d} f(t)\,\overline{g(t-x)}\,e^{-2\pi i t\cdot\omega}\,dt \quad \text{for } x,\omega \in \mathbb{R}^d. \tag{29}$$

For $f, g \in \mathrm{L}^2(\mathbb{R}^d)$, the STFT $\mathcal{V}_g f$ is uniformly continuous (by Riemann–Lebesgue) on $\mathbb{R}^{2d}$ and can be written as

$$(\mathcal{V}_g f)(x,\omega) = \widehat{f\cdot T_x\bar{g}}(\omega) \tag{30}$$

$$= \langle f, M_\omega T_x g\rangle_{\mathrm{L}^2} \tag{31}$$

$$= e^{-2\pi i x\cdot\omega}(f * M_\omega g^*)(x). \tag{32}$$

The STFT as a function in x and ω seems to provide the possibility to obtain information about the occurrence of arbitrary frequencies ω at arbitrary locations x as desired. However, the *uncertainty principle* (cf. [51]) implies that there is a limitation concerning the joint resolution. In fact, the STFT has limitations in its time–frequency resolution capability: Low frequencies can hardly be located with narrow windows, and similarly, short pulses remain invisible for wide windows. The choice of the analyzing window is therefore crucial.

Just like the Fourier transform, the STFT is a kind of time–frequency representation of a signal. This again raises the question of how to reconstruct the signal from its time–frequency representation. To approach this, we need the orthogonality relations of the STFT, which corresponds to Parseval's formula (20) for the Fourier transform:

Theorem 7 (Orthogonality relations for STFT). *Let* $f_1, f_2, g_1, g_2 \in \mathrm{L}^2(\mathbb{R}^d)$. *Then* $\mathcal{V}_{g_j} f_j \in \mathrm{L}^2(\mathbb{R}^{2d})$ *for* $j \in \{1,2\}$, *and*

$$\langle \mathcal{V}_{g_1} f_1, \mathcal{V}_{g_2} f_2\rangle_{\mathrm{L}^2(\mathbb{R}^{2d})} = \langle f_1, f_2\rangle_{\mathrm{L}^2}\overline{\langle g_1, g_2\rangle_{\mathrm{L}^2}}.$$

Corollary 3. *If* $f, g \in \mathrm{L}^2(\mathbb{R}^d)$, *then*

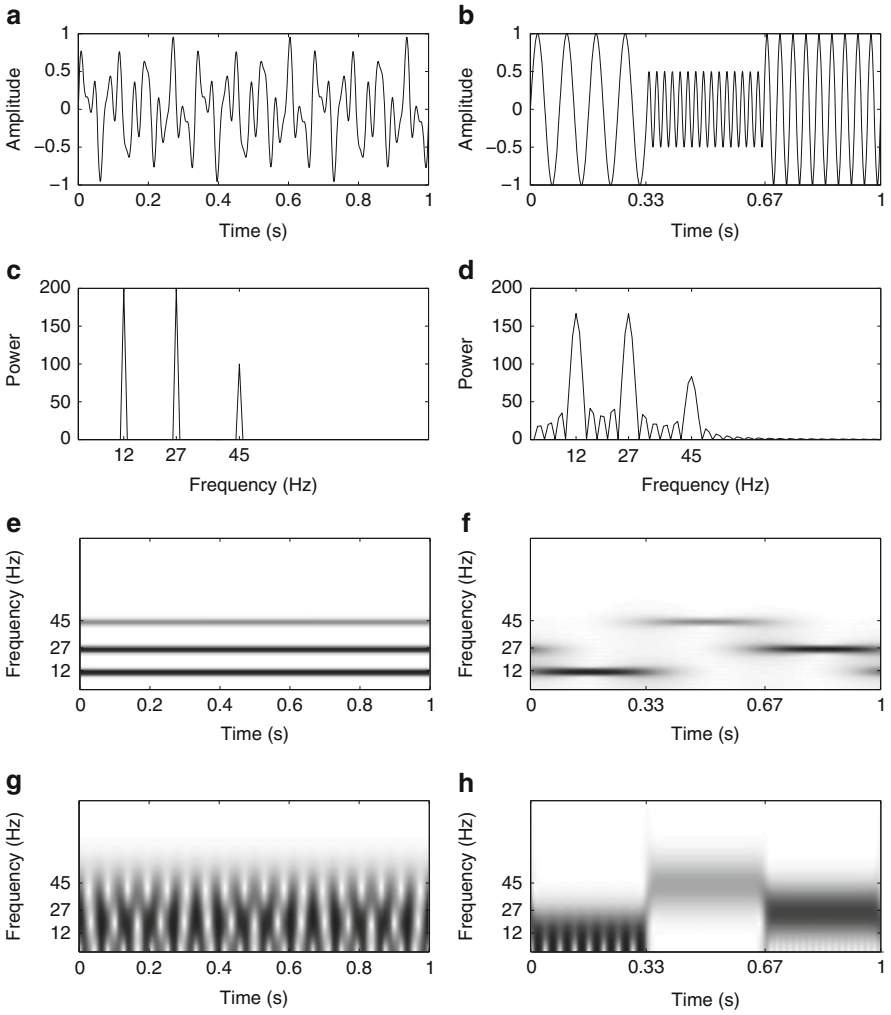


Fig. 1 Two signals and their (short-time) Fourier transforms. (**a**) Signal 1: Concurrent frequencies. (**b**) Signal 2: Consecutive frequencies. (**c**) Fourier power spectrum 1. (**d**) Fourier power spectrum 2. (**e**) STFT 1 with wide window. (**f**) STFT 2 with wide window. (**g**) STFT 1 with narrow window. (**h**) STFT 2 with narrow window

$$\|\mathcal{V}_g f\|_{\mathrm{L}^2(\mathbb{R}^{2d})} = \|f\|_{\mathrm{L}^2}\|g\|_{\mathrm{L}^2}.$$

In the case of $\|g\|_{\mathrm{L}^2} = 1$, *we have*

$$\|f\|_{\mathrm{L}^2} = \|\mathcal{V}_g f\|_{\mathrm{L}^2(\mathbb{R}^{2d})} \quad \forall f \in \mathrm{L}^2(\mathbb{R}^d), \tag{33}$$

i.e., the STFT as an isometry from $\mathrm{L}^2(\mathbb{R}^d)$ *into* $\mathrm{L}^2(\mathbb{R}^{2d})$.

Formula (33) shows that the STFT preserves the energy of a signal; it corresponds to (19) which shows the same property for the Fourier transform. Therefore, f is completely determined by $\mathcal{V}_g f$, and the inversion is given by a vector-valued integral (for good functions valid in the pointwise sense):

Corollary 4 (Inversion formula for the STFT). *Let $g, \gamma \in \mathrm{L}^2(\mathbb{R}^d)$ and $\langle g, \gamma \rangle \neq 0$. Then*

$$f(x) = \frac{1}{\langle \gamma, g \rangle_{\mathrm{L}^2}} \iint_{\mathbb{R}^{2d}} \mathcal{V}_g f(x, \omega)\ M_\omega T_x \gamma(x)\ d\omega\, dx \quad \forall f \in \mathrm{L}^2(\mathbb{R}^d). \tag{34}$$

Obviously, $\gamma = g$ is a natural choice here. The time–frequency analysis of signals is usually done by three subsequent steps:

(i) *Analysis*: Using the STFT, the signal is transformed into a joint time–frequency representation.
(ii) *Processing*: The obtained signal representation is then manipulated in a certain way, e.g., by restriction to a part of the signal yielding the relevant information.
(iii) *Synthesis*: The inverse STFT is applied to the processed representation, thus creating a new signal.

A function is completely represented by its STFT but in a highly redundant way. To minimize the influence of the uncertainty principle, the analyzing window g should be chosen such that g and its Fourier transform $\hat{g}$ both decay rapidly, e.g., as Schwartz functions. A computational implementation can only be obtained by a discretization of both the functions and the STFT. Therefore, only sampled versions of the STFT are possible, and only certain locations and frequencies are used for analyzing a given signal. The challenge is to find the appropriate lattice constants in time and frequency and to obtain good time–frequency resolution.

4 Gabor Frames in $\mathrm{L}^2(\mathbb{R}d)$

By formula (31), the STFT analyzes a function $f \in \mathrm{L}^2(\mathbb{R}^d)$ into coefficients $\langle f, M_\omega T_x g \rangle_{\mathrm{L}^2}$ using modulations and translations of a single window function $g \in \mathrm{L}^2(\mathbb{R}^d) \backslash \{0\}$. One problem we noticed was that these TF-shifts are infinitesimal and overlap largely, making the STFT a highly redundant time–frequency representation. An idea to overcome this is to restrict to discrete choices of time positions x and frequencies ω such that this redundancy is decreased while leaving enough information in the coefficients about the time–frequency behavior of f. This is the very essence of Gabor analysis: It is sought to expand functions in $\mathrm{L}^2(\mathbb{R}^d)$ into an absolutely convergent series of modulations and translations of a window function g. Therefore, it is interesting to find necessary and sufficient conditions on g and a discrete set $\Lambda \subseteq \mathbb{R}^d \times \mathbb{R}^d$ such that

$$\{g_{x,\omega}\}_{(x,\omega)\in\Lambda} := \{M_\omega T_x g\}_{(x,\omega)\in\Lambda}$$

forms a frame for $\mathrm{L}^2(\mathbb{R}^d)$. The question arises how the sampling set Λ should be structured. It turns out to be very convenient to have this set closed under the addition operation, urging Λ to be a subgroup of the time–frequency plane, i.e., $\Lambda \trianglelefteq \mathbb{R}^d \times \mathbb{R}^d$. Dennis Gabor (actually *Dénes Gábor*) suggested in his *Theory of Communication* [45], 1946, to use fixed step sizes $\alpha, \beta > 0$ for time and frequency and use the set $\{\alpha k\}_{k\in\mathbb{Z}^d}$ for the time positions and $\{\beta n\}_{n\in\mathbb{Z}^d}$ for the frequencies, yielding the functions

$$g_{k,n}(x) := M_{\beta n} T_{\alpha k} g(x) = e^{2\pi \mathrm{i} \beta n \cdot x} g(x - \alpha k)$$

as analyzing elements. This is the approach that is usually presented in the literature, although there is also a more general group theoretical setting possible where Λ is an arbitrary (discrete) subgroup. This subgroup is also called a *time–frequency lattice*, although it does not have to be of such a "rectangular" shape in general.

Definition 10. A *lattice* $\Lambda \subseteq \mathbb{R}^d$ is a (discrete) subgroup of $\mathbb{R}^d$ of the form $\Lambda = \mathfrak{A}\mathbb{Z}^d$, where $\mathfrak{A}$ is an invertible $d \times d$-matrix over $\mathbb{R}$. Lattices in $\mathbb{R}^{2d}$ can be described as

$$\Lambda = \left\{(x, y) \in \mathbb{R}^{2d} \,\middle|\, (x, y) = (Ak + B\ell, Ck + D\ell),\ (k, \ell) \in \mathbb{Z}^{2d}\right\}$$

with $A, B, C, D \in \mathbb{C}^{d\times d}$ and

$$\mathfrak{A} = \begin{pmatrix} A & B \\ C & D \end{pmatrix}.$$

A lattice $\Lambda = \alpha\mathbb{Z}^d \times \beta\mathbb{Z}^d \trianglelefteq \mathbb{R}^{2d}$ for $\alpha, \beta > 0$ is called a *separable lattice,* a *product lattice,* or a *grid*.

In the following, our lattice will be of the separable type for fixed lattice parameters $\alpha, \beta > 0$.

Definition 11. For a nonzero window function $g \in \mathrm{L}^2(\mathbb{R}^d)$ and lattice parameters $\alpha, \beta > 0$, the set of time–frequency shifts

$$\mathcal{G}(g, \alpha, \beta) := \{M_{\beta n} T_{\alpha k} g\}_{k,n\in\mathbb{Z}^d}$$

is called a *Gabor system*. If $\mathcal{G}(g, \alpha, \beta)$ is a frame for $\mathrm{L}^2(\mathbb{R}^d)$, it is called a *Gabor frame* or *Weyl–Heisenberg frame*. The associated frame operator is the *Gabor frame operator* and takes the form

$$Sf = \sum_{k,n\in\mathbb{Z}^d} \sum \langle f, M_{\beta n} T_{\alpha k} g\rangle_{\mathrm{L}^2}\, M_{\beta n} T_{\alpha k} g \tag{35}$$

$$= \sum_{k,n\in\mathbb{Z}^d} \sum \mathcal{V}_g f(\alpha k, \beta n)\, M_{\beta n} T_{\alpha k} g$$

for all $f \in \mathrm{L}^2(\mathbb{R}^d)$. The window g is also called the *Gabor atom*.

According to the general frame theory, $\{S^{-1} g_{k,n}\}_{k,n\in\mathbb{Z}^d}$ yields the *canonical dual frame*. So we would have to compute S^{-1} and apply it to *all* modulated and translated versions of the Gabor atom g. A direct computation shows that for arbitrary fixed indices $\ell, m \in \mathbb{Z}^d$,

$$S M_{\beta m} T_{\alpha \ell} = M_{\beta m} T_{\alpha \ell} S. \tag{36}$$

Consequently, also S^{-1} commutes with time–frequency shifts, which gives the following fundamental result for (regular) Gabor analysis:

Theorem 8. *If the given Gabor system $\mathcal{G}(g, \alpha, \beta)$ is a frame for $\mathrm{L}^2(\mathbb{R}^d)$, then all of the following hold:*

(a) *There exists a dual window $\gamma \in \mathrm{L}^2(\mathbb{R}^d)$ such that the dual frame is given by the Gabor frame $\mathcal{G}(\gamma, \alpha, \beta)$.*
(b) *Every $f \in \mathrm{L}^2(\mathbb{R}^d)$ has an expansion of the form*

$$f = \sum_{k,n\in\mathbb{Z}^d} \sum \langle f, M_{\beta n} T_{\alpha k} g\rangle_{\mathrm{L}^2}\, M_{\beta n} T_{\alpha k} \gamma \tag{37}$$

$$= \sum_{k,n\in\mathbb{Z}^d} \sum \langle f, M_{\beta n} T_{\alpha k} \gamma\rangle_{\mathrm{L}^2}\, M_{\beta n} T_{\alpha k} g$$

with unconditional convergence in $\mathrm{L}^2(\mathbb{R}^d)$.
(c) *The canonical dual frame is given by the Gabor frame $\{M_{\beta n} T_{\alpha k} S^{-1} g\}_{k,n\in\mathbb{Z}^d}$ built from the canonical dual window $\gamma^\circ := S^{-1} g$.*
(d) *The inverse frame operator S^{-1} is just the frame operator for the Gabor system $\mathcal{G}(\gamma^\circ, \alpha, \beta)$ and*

$$S^{-1} f = \sum_{k,n\in\mathbb{Z}^d} \sum \left\langle f, M_{\beta n} T_{\alpha k} \gamma^\circ \right\rangle_{\mathrm{L}^2}\, M_{\beta n} T_{\alpha k} \gamma^\circ. \tag{38}$$

We note that if the function g is compactly supported and the modulation parameter β is sufficiently small, it is easy to verify whether $\mathcal{G}(g, \alpha, \beta)$ is a frame and to find the canonical dual window in the affirmative case; see [51, 6.4] or [17, 9.1].

One can show [51, 7.6.1] that all dual windows γ of a Gabor frame $\mathcal{G}(g, \alpha, \beta)$ are within an affine subspace of $\mathrm{L}^2(\mathbb{R}^d)$, namely, $\gamma \in \gamma^\circ + \mathcal{K}^\perp$, where $\mathcal{K}$ is the closed linear span of $\mathcal{G}\left(g, \frac{1}{\beta}, \frac{1}{\alpha}\right)$ and therefore

$$\mathcal{K}^{\perp} = \left\{ h \in \mathrm{L}^2(\mathbb{R}^d) : \langle h, M_{n/\alpha} T_{k/\beta} g \rangle_{\mathrm{L}^2} = 0 \ \ \forall k, n \in \mathbb{Z}^d \right\}. \tag{39}$$

Hence, we have $\gamma = \gamma^\circ + h$ for a certain $h \in \mathcal{K}^{\perp}$, and as $\gamma^\circ \in \mathcal{K}$, the canonical dual window possesses the smallest L^2-norm among all dual windows and is most similar to the original window g. However, there might be reasons not to choose the canonical dual window, but one of the others in $\gamma^\circ + \mathcal{K}^{\perp}$, if, e.g., one wants the dual window to have a smaller essential support or if the window should be as smooth as possible. Explicit constructions of alternative dual windows can be found in [17].

A key result in Gabor analysis states a necessary condition for a Gabor system to form a frame:

Theorem 9. *Let $g \in \mathrm{L}^2(\mathbb{R}^d) \backslash \{0\}$ and $\alpha, \beta > 0$. If $\mathcal{G}(g, \alpha, \beta)$ is a frame, then:*

(a) $\alpha\beta \leq 1$.
(b) *$\mathcal{G}(g, \alpha, \beta)$ is a basis if and only if $\alpha\beta = 1$.*

Unfortunately, having $\alpha\beta \leq 1$ is not sufficient for a Gabor system to form a frame. Sufficient conditions are presented, e.g., in [16, 8.4]. A special result is known for the Gaussian function:

Theorem 10. *Consider the normalized Gaussian $\varphi(x) := 2^{d/4} e^{-\pi x^2}$. Then $\mathcal{G}(\varphi, \alpha, \beta)$ is a frame for $\mathrm{L}^2(\mathbb{R}^d)$ if and only if $\alpha\beta < 1$.*

In signal analysis, it is customary to call the case

$$\begin{aligned} \alpha\beta &< 1 \ \textit{oversampling} \\ \alpha\beta &= 1 \ \textit{critical sampling} \\ \alpha\beta &> 1 \ \textit{undersampling} \end{aligned}$$

In the case of the Gaussian window, oversampling guarantees an excellent time–frequency localization. But for Gabor frame theory in $\mathrm{L}^2(\mathbb{R}^d)$, it is quite delicate to find appropriate windows for given $\alpha\beta \leq 1$. The case $\alpha\beta = 1$ is problematic from the point of view of time–frequency analysis, as the Balian–Low theorem demonstrates:

Theorem 11 (Balian–Low). *Let $g \in \mathrm{L}^2(\mathbb{R}^d)$ be a nonzero window and $\alpha, \beta > 0$ with $\alpha\beta = 1$. If g has good TF-concentration in the sense of*

$$\|Xg\|_{\mathrm{L}^2} \|X\hat{g}\|_{\mathrm{L}^2} < \infty,$$

then $\mathcal{G}(g, \alpha, \beta)$ cannot constitute a frame.

Combining Theorems 9 and 10 shows that it is impossible for a Gabor basis to be well localized in both the time domain and the frequency domain. This motivates the study of redundant Gabor systems: As demonstrated by Theorem 10, redundant Gabor frames exist for any $\alpha\beta < 1$.

5 Discrete Gabor Systems

For practical implementations of Gabor analysis, it is essential to develop discrete versions of the theory for Gabor frames.

Gabor Frames in $\ell^2(\mathbb{Z})$

Classically, most signals were considered as "continuous waves." Indeed, the technology for signal processing originally was of the continuous-time analog type before digital computers came into our everyday life. Nowadays, digital signal processing is used almost exclusively, forcing us to change our function model to a time-discrete one. It is therefore natural to switch from $L^2(\mathbb{R})$ to $\ell^2(\mathbb{Z})$.

Gabor frame theory in $\ell^2(\mathbb{Z})$ is very similar to that in $L^2(\mathbb{R})$ and will therefore only be discussed briefly in this section. The main differences concern the time shifts and frequency shifts. Time shifts are given as multiples of integer translates, i.e.,

$$T_k f(j) = f(j-k) \tag{40}$$

for $k \in \mathbb{Z}$ and $f \in \ell^2(\mathbb{Z})$. A shift parameter $\alpha > 0$ for Gabor frames in $\ell^2(\mathbb{Z})$ can only be given as $\alpha = N \in \mathbb{N}$.

For fixed $L \in \mathbb{N}$ and corresponding to the modulation parameter $1/L$, we define the modulation operator M_ℓ by

$$M_\ell f(j) = e^{2\pi \mathrm{i} j\ell/L} f(j) \tag{41}$$

for $\ell \in \mathbb{Z}$. Modulations are now periodic with period L, i.e., $M_{\ell+nL} = M_\ell \quad \forall n \in \mathbb{Z}$, implying that one needs only the modulations $M_0, \ldots, M_{L-1}$.

The *discrete Gabor system* generated by the sequence $g \in \ell^2(\mathbb{Z})$, shift parameters N, and modulation parameter $1/L$ is now the family of sequences $\{g_{k,\ell}\}_{k\in\mathbb{Z},\ell\in\langle L\rangle}$ where

$$g_{k,\ell}(j) := M_\ell T_{kN} g(j) = e^{2\pi \mathrm{i} j\ell/L} g(j-kN)$$

and $\langle L\rangle := \{0, \ldots, L-1\} \subseteq \mathbb{Z}$.

If a Gabor system satisfies the frame inequalities for $f \in \ell^2(\mathbb{Z})$, the dual frame is again a Gabor frame built from a dual window $\gamma \in \ell^2(\mathbb{Z})$. The frame expansion takes the form

$$f = \sum_{k=-\infty}^{\infty} \sum_{\ell=0}^{L-1} \langle f, M_\ell T_{kN} \gamma \rangle_2 \, M_\ell T_{kN} g \quad \text{for} \quad f \in \ell^2(\mathbb{Z}).$$

Many results and conditions for Gabor systems in $\ell^2(\mathbb{Z})$ can *mutatis mutandis* be taken over from $\mathrm{L}^2(\mathbb{R})$, e.g., a necessary condition for the mentioned Gabor system to be a frame for $\ell^2(\mathbb{Z})$ is that $\alpha\beta = N/L \leq 1$.

We note that there is a natural way of constructing Gabor frames in $\ell^2(\mathbb{Z})$ from Gabor frames in $L^2(\mathbb{R})$ through sampling; see the paper [55] by Janssen.

The step from $\mathrm{L}^2(\mathbb{R})$ to $\ell^2(\mathbb{Z})$ is the first one toward computational realization of Gabor analysis. However, since in finite time only finitely many elements can be considered, only vectors of finite length and finite sums can be computed. Therefore, we turn to signals of finite length next.

Finite Discrete Periodic Signals

In practice, one has to resort to finite, discrete sequences. We will consider signals $f \in \mathbb{C}^L$, i.e., signals of length $L \in \mathbb{N}$, and write $f = (f(0), \ldots, f(L-1))$, defined (for convenience) over the domain $\langle L \rangle := \{0, \ldots, L-1\} \subseteq \mathbb{Z}$. This way of indexing suggests in a natural way to view them as functions over the group of unit roots of order L or equivalently as periodic sequences with

$$f(j + nL) := f(j) \quad \forall n \in \mathbb{Z}, j \in \langle L \rangle.$$

The discrete modulation M_ℓ defined in (41) can still be applied; the translation T_k defined in (40) can be taken from the range $0 \leq k \leq L-1$.

The *discrete Fourier transform* (DFT) of $f \in \mathbb{C}^L$ is defined as

$$\hat{f}(j) := (\mathcal{F} f)(j) := \sum_{k=0}^{L-1} f(k) \, e^{-2\pi \mathrm{i} jk/L}, \quad j \in \mathbb{Z}_L, \tag{42}$$

which is – up to a constant – a unitary mapping on $\mathbb{C}^L$. Its inverse is

$$(\mathcal{F}^{-1} f)(j) := \frac{1}{L} \sum_{k=0}^{L-1} f(k) \, e^{2\pi \mathrm{i} jk/L}, \quad j \in \mathbb{Z}_L. \tag{43}$$

The unitary version $\mathbb{C}^L \to \mathbb{C}^L$ has the factor $1/\sqrt{L}$ in front. A well-known and very efficient implementation of the DFT is the *fast Fourier transform* (FFT).

The discrete STFT of $f \in \mathbb{C}^L$ with respect to the discrete window $g \in \mathbb{C}^L$ is given as

$$(\mathcal{V}_g f)(k, \ell) = \langle f, M_\ell T_k g \rangle_{\mathbb{C}^L}.$$

The actions of time and frequency shifts are in more detail given as

$$T_k f = T_k\big(f(0),\dots,f(L-1)\big) = \big(f(-k),\dots,f(L-1-k)\big)$$

and

$$\begin{aligned} M_\ell f &= M_\ell\big(f(0),\dots,f(L-1)\big) \\ &= \big(f(0), e^{2\pi \mathrm{i}\ell/L} f(1), e^{2\pi \mathrm{i}2\ell/L} f(2),\dots, e^{2\pi \mathrm{i}(L-1)\ell/L} f(L-1)\big). \end{aligned}$$

The actions of the TF-shifts can be described as matrices that operate on the vector $f = (f(0),\dots,f(L-1))^{\mathrm{T}}$. The time-shift matrix T_k is given as the permutation matrix with ones on the (periodized) k-th subdiagonal, whereas the modulation matrix has its exponential entries positioned at the main diagonal. It is obvious that the composition of arbitrary TF-shifts need not be commutative, since

$$T_k M_\ell = e^{2\pi \mathrm{i}k\ell/L} M_\ell T_k, \quad k,\ell \in \mathbb{Z}_L$$

To get a more compact notation for TF-shifts, we write

$$\pi(\lambda) := \pi(k,\ell) := M_\ell T_k \quad \text{with} \quad \lambda = (k,\ell) \in \mathbb{Z}_L \times \mathbb{Z}_L,$$

where $\mathbb{Z}_L \times \mathbb{Z}_L$ is the discrete time–frequency plane. The commutation relations imply for $\lambda = (r,m)$ and $\mu = (s,n)$

$$\pi(\lambda)\,\pi(\mu) = \pi(\lambda+\mu)\, e^{2\pi \mathrm{i}rn/L} \tag{44}$$

$$= \pi(\mu)\,\pi(\lambda)\, e^{2\pi \mathrm{i}(rn-sm)/L}. \tag{45}$$

Frames and Gabor Frames in $\mathbb{C}^L$

The general frame definitions and results can easily be carried over to the case of finite discrete signals. The conditions for the finite sequence $\{g_0,\dots,g_{N-1}\}$ of elements $g_j \in \mathbb{C}^L$ to be a frame for the finite-dimensional Hilbert space $\mathbb{C}^L$ are that there exist $A, B > 0$ such that

$$A\sum_{k=0}^{L-1} |f(k)|^2 \le \sum_{j=0}^{N-1} \big|\langle f, g_j\rangle_{\mathbb{C}^L}\big|^2 \le B\sum_{k=0}^{L-1} |f(k)|^2 \quad \forall f \in \mathbb{C}^L$$

or

$$A\|f\|_2^2 \le \|Cf\|_2^2 \le B\|f\|_2^2 \quad \forall f \in \mathbb{C}^L,$$

where C is the analysis operator. It is obvious that the sequence $\{g_j\}_{j=1}^{N-1}$ has to span all of $\mathbb{C}^L$, i.e., $span\{g_j\}_{j=0}^{N-1} = \mathbb{C}^L$; hence, $N \geq L$ in a Hilbert space with dimension L. Also the converse is true: Every spanning set in $\mathbb{C}^L$ is a frame for $\mathbb{C}^L$.

The action of the linear analysis operator C on the vector f is given as the vector $Cf = \left(\langle f, g_j\rangle\right)_{j=1}^{N-1}$, indicating that its j-th entry is

$$(Cf)_j = \langle f, g_j\rangle = \sum_{k=0}^{L-1} f(k)\,\overline{g_j(k)}\,.$$

Letting $g^* = \bar{g}^{\mathrm{T}}$, the matrix form of $C \in \mathbb{C}^{N\times L}$ is

$$C = \begin{pmatrix} g_0^* \\ \vdots \\ g_{N-1}^* \end{pmatrix} = \begin{pmatrix} \overline{g_0(0)} & \cdots & \overline{g_0(L-1)} \\ \vdots & \vdots & \vdots \\ \overline{g_{N-1}(0)} & \cdots & \overline{g_{N-1}(L-1)} \end{pmatrix}.$$

A family $\{g_j\}_{j\in\langle N\rangle}$ is a frame for $\mathbb{C}^L$ if and only if the corresponding analysis operator C has full rank, and every matrix with full rank uniquely represents a frame.

The frame operator $S = C^*C$ becomes an $L \times L$ matrix that also has full rank, and it is therefore invertible. Its condition number equals the ratio between its largest and smallest eigenvalue; letting A denote the largest lower frame bound and B the smallest upper frame bound, this is equal to the ratio B/A.

If we translate the discrete frame expansion

$$f = C^*c = (g_0, \ldots, g_{N-1}) \begin{pmatrix} c(0) \\ \vdots \\ c(N-1) \end{pmatrix} = \begin{pmatrix} \sum_{j=0}^{N-1} c(j)\, g_j(0) \\ \vdots \\ \sum_{j=0}^{N-1} c(j)\, g_j(L-1) \end{pmatrix}$$

for a given $f \in \mathbb{C}^L$, we see from a linear algebra point of view that we are looking for N unknown coordinates of $c \in \mathbb{C}^N$, using $L \leq N$ equations. Clearly, the solution cannot be unique if $L < N$. Considering that

$$f = SS^{-1}f = C^*C(C^*C)^{-1}f,$$

we see that one solution for c could be given as

$$c = C(C^*C)^{-1}f = (C^*)^{\dagger}f$$

in terms of the pseudo-inverse of the synthesis operator C^*. This also provides the matrix form of the canonical dual frame that is given by

$$\left(S^{-1}g_0, \ldots, S^{-1}g_{N-1}\right)^* = \left(S^{-1}C^*\right)^* = CS^{-1} = (C^*)^{\dagger}.$$

We will now proceed to the special case of Gabor frames. They are given as a sequence of TF-shifts of a single window function $g \in \mathbb{C}^L$, i.e., a Gabor frame for $\mathbb{C}^L$ is a sequence $\{g_\lambda\}_{\lambda\in\Lambda} := \{\pi(\lambda)g\}_{\lambda\in\Lambda}$ for a certain discrete subset $\Lambda \subseteq \mathbb{Z}_L \times \mathbb{Z}_L$. We write C_g for the Gabor analysis operator to indicate the dependence on g and use it synonymously for the Gabor frame itself. It is clear that it is necessary to have $N \geq L$ elements to span all of $\mathbb{C}^L$, but this is of course not sufficient for validating a frame. The ratio between N and L is also called the *redundancy* of the frame,

$$\mathrm{red}_C := \frac{N}{L}.$$

For any subgroup $\Lambda \trianglelefteq \mathbb{Z}_L \times \mathbb{Z}_L$, the Gabor frame operator $S_g = C_g^* C_g$ commutes with all TF-shifts $\pi(\lambda)$ for $\lambda \in \Lambda$. This can be shown in a similar way as in Sect. 4. Therefore, the dual frame is once again a Gabor frame, built by the same TF-shifts of a single dual window $\gamma \in \mathbb{C}^L$. The canonical dual frame consists of elements

$$S_g^{-1}\pi(\lambda)g = \pi(\lambda)S_g^{-1}g = \pi(\lambda)\gamma^\circ,$$

and the computation of the canonical dual window reduces to finding a solution for the linear equation

$$S_g\gamma^\circ = g. \tag{46}$$

Therefore, the discrete Gabor expansion of an $f \in \mathbb{C}^L$ is given as

$$f = \sum_{\lambda\in\Lambda} \langle f, \pi(\lambda)g\rangle_{\mathbb{C}^L}\, \pi(\lambda)\gamma^\circ = \sum_{\lambda\in\Lambda} \langle f, \pi(\lambda)\gamma^\circ\rangle_{\mathbb{C}^L}\, \pi(\lambda)g,$$

where the Gabor coefficients belong to $\ell^2(\Lambda) \cong \mathbb{C}^N$.

A special case for a lattice is a so-called separable lattice $\Lambda = \alpha\mathbb{Z}_L \times \beta\mathbb{Z}_L$ with $\alpha, \beta \in \mathbb{N}$ being divisors of L. The elements of such a Gabor frame take the form

$$M_{\beta\ell}T_{\alpha k}g(j) = e^{2\pi \mathrm{i}\beta\ell j/L}g(j-\alpha k)$$

with $k \in \left\langle \frac{L}{\alpha} \right\rangle$ and $\ell \in \left\langle \frac{L}{\beta} \right\rangle$. The number of elements is $N = \frac{L}{\alpha}\cdot\frac{L}{\beta} = \frac{L^2}{\alpha\beta}$, and it is necessary to have $\frac{L^2}{\alpha\beta} \geq L$ elements to have a frame. The oversampled case is therefore given for $\alpha\beta < L$, and the undersampled case for $\alpha\beta > L$. Critical sampling is given for $\alpha\beta = L$.

6 Image Representation by Gabor Expansion

We have seen that Gabor analysis can be considered as a localized Fourier analysis, where the main design freedom is the choice of (a) the time–frequency lattice and (b) the window function. The type of sampling lattice can be distinguished into a separable or non-separable case, where the first one can be described by the choice of lattice constants $\alpha, \beta > 0$.

It turns out that in the twofold-separable case, i.e., where the d-dimensional analysis window is a tensor product of d one-dimensional functions

$$\boldsymbol{g} = g_1 \otimes \cdots \otimes g_d\,, \quad \text{with} \quad g_1 \otimes \cdots \otimes g_d(x_1, \ldots, x_d) = g_1(x_1) \ldots g_d(x_d),$$

and the sampling lattice Λ is a product $\Lambda = \prod_{i=1}^{d} \alpha_i \mathbb{Z}_{L_i} \times \prod_{i=1}^{d} \beta_i \mathbb{Z}_{L_i}$, the dual Gabor window $\boldsymbol{\gamma}$ is given as a product $\boldsymbol{\gamma} = \gamma_1 \otimes \cdots \otimes \gamma_d$ as well. Thus, the computation is reduced to finding the 1D duals γ_i of the 1D atoms g_i with respect to the corresponding 2D time–frequency lattices $\Lambda_i = \alpha_i \mathbb{Z}_{L_i} \times \beta_i \mathbb{Z}_{L_i}$.

Our aim here is to show how the results can be applied to the case of image signals. Gabor expansions of finite discrete 2D signals (i.e., digital images) are similar to those of finite discrete 1D signals, and in a more general notation, there is no difference at all. We are going to describe it next (Fig. 2).

2D Gabor Expansions

The key point for the development of efficient algorithms is to interpret an image of size $L_1 \times L_2$ as a real- or complex-valued function on the additive Abelian group $\mathcal{G} = \mathbb{Z}_{L_1} \times \mathbb{Z}_{L_2}$. The position–frequency space is

$$\mathcal{G} \times \widehat{\mathcal{G}} = \mathbb{Z}_{L_1} \times \mathbb{Z}_{L_2} \times \widehat{\mathbb{Z}_{L_1} \times \mathbb{Z}_{L_2}}.$$

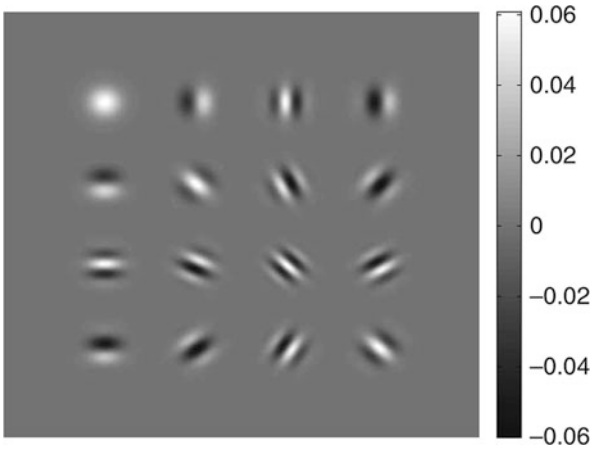


Fig. 2 Typical 2D Gabor atoms

A Gabor system $\mathcal{G}(\boldsymbol{g}, \Lambda)$ consists of TF-shifts $M_l T_k \boldsymbol{g}$ of a window $\boldsymbol{g} \in \mathbb{C}^{L_1 \times L_2}$, where $(\boldsymbol{k}, \boldsymbol{l})$ are elements of a sampling subgroup $\Lambda \trianglelefteq \mathbb{Z}_{L_1} \times \mathbb{Z}_{L_2} \times \widehat{\mathbb{Z}_{L_1} \times \mathbb{Z}_{L_2}}$. The Gabor coefficients of the image $\boldsymbol{f} \in \mathbb{C}^{L_1 \times L_2}$ are defined as

$$c_{\boldsymbol{k},\boldsymbol{l}} := \langle \boldsymbol{f}, M_l T_k \boldsymbol{g} \rangle_{\mathrm{F}}, \quad (\boldsymbol{k}, \boldsymbol{l}) \in \Lambda. \tag{47}$$

Here, we use the subscript F in order to recall that for matrices (this is how images are usually stored), one takes the scalar product and the corresponding norm just as the Euclidian one in $\mathbb{C}^N$, with $N = L_1 L_2$, usually denoted as Frobenius norm.

The Gabor system is a frame if for $0 < A \leq B < \infty$ one has

$$A \|\boldsymbol{f}\|_{\mathrm{F}}^2 \leq \sum_{(\boldsymbol{k},\boldsymbol{l}) \in \Lambda} |\langle \boldsymbol{f}, M_l T_k \boldsymbol{g} \rangle_{\mathrm{F}}|^2 \leq B \|\boldsymbol{f}\|_{\mathrm{F}}^2 \quad \forall \boldsymbol{f} \in \mathbb{C}^{L_1 \times L_2}.$$

For dimensionality reasons, it is clear that the frame condition is only possible if the number of elements in Λ has to be at least equal to the dimension of the signal space, and, therefore, we need $L_1 L_2 \leq |\Lambda| \leq (L_1 L_2)^2$. The redundancy of the Gabor frame is

$$\mathrm{red}_\Lambda := \frac{|\Lambda|}{L_1 L_2} \geq 1.$$

As in the one-dimensional case, the Gabor frame operator

$$S_{\boldsymbol{g}} \boldsymbol{f} := \sum_{(\boldsymbol{k},\boldsymbol{l}) \in \Lambda} \langle \boldsymbol{f}, M_l T_k \boldsymbol{g} \rangle_{\mathrm{F}} \, M_l T_k \boldsymbol{g}$$

commutes with TF-shifts determined by Λ, and the minimal resp. maximal eigenvalue are equal to the maximal lower frame bound A and minimal upper frame bound B, respectively.

Again, the dual Gabor frame has a similar structure as the Gabor frame itself: using the same TF-shifts, now applied to a dual window $\boldsymbol{\gamma} \in \mathbb{C}^{L_1 \times L_2}$, one has the expansion

$$\boldsymbol{f} = \sum_{(\boldsymbol{k},\boldsymbol{l}) \in \Lambda} \langle \boldsymbol{f}, M_l T_k \boldsymbol{g} \rangle_{\mathrm{F}} \, M_l T_k \boldsymbol{\gamma} = \sum_{(\boldsymbol{k},\boldsymbol{l}) \in \Lambda} \langle \boldsymbol{f}, M_l T_k \boldsymbol{\gamma} \rangle_{\mathrm{F}} \, M_l T_k \boldsymbol{g}$$

for all $\boldsymbol{f} \in \mathbb{C}^{L_1 \times L_2}$. The existence of the dual atom is guaranteed by the theory of frames, and the calculation of the dual Gabor frame is done by the methods developed there. Recent results guarantee that good TF-concentration of the atom g implies a similar quality for the dual Gabor atom. Typically, the condition number of the frame operator depends on the geometric density (hence, to some extent, on the redundancy) of the lattice. However it is worth mentioning that even for low redundancy factors, relatively good condition numbers can be expected for

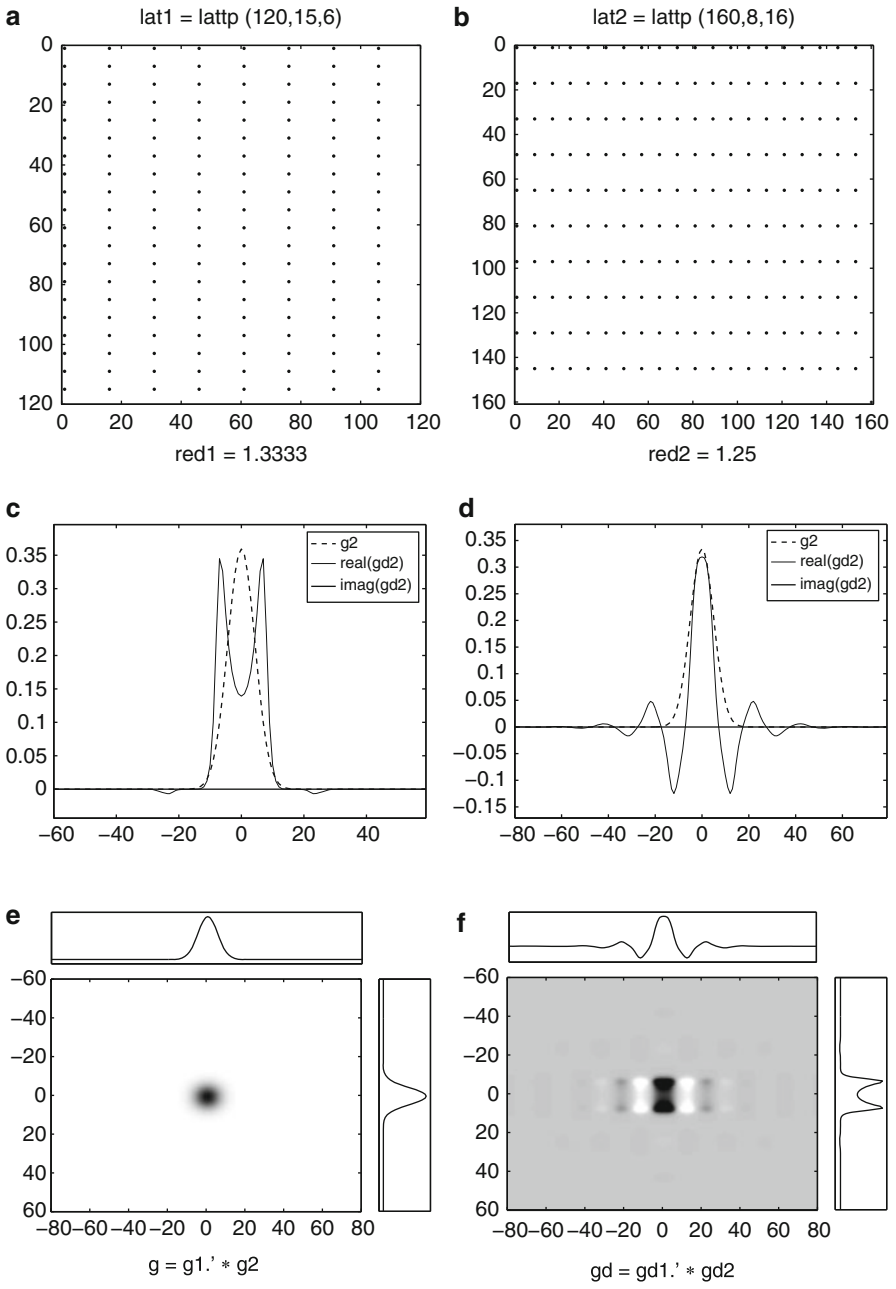


Fig. 3 2D separable window and its dual on a fully separable lattice. (**a**) Lattice $\Lambda_1 = 15\,\mathbb{Z}_{120} \times 6\mathbb{Z}_{120}$. (**b**) Lattice $\Lambda_2 = 8\,\mathbb{Z}_{160} \times 16\mathbb{Z}_{160}$. (**c**) Guassian g1 and its dual γ_1° on Λ_1. (**d**) Guassian g2 and its dual γ_2° on Λ_2. (**e**) 2D window g = g1 $\otimes$ g2. (**f**) 2D dual window $\gamma^\circ = \gamma_1^\circ \otimes \gamma_2^\circ$ on $\Lambda_1 \times \Lambda_2$

suitably chosen atoms and that perfect reconstruction can be achieved in a stable way in a computationally efficient way even if the discretization of the continuous representation formula is far from satisfactory. Expressed differently, the frame operator may be far away from the identity operator but still stably invertible.

The optimal method and effective computational cost for obtaining Gabor expansions of an image depend on the structure of the 4D sampling lattice. A (fully) separable position–frequency lattice (PF-lattice) can be described by parameters $\alpha_1, \alpha_2, \beta_1, \beta_2 > 0$ such that the constants α_i and β_i describing the position and frequency shift parameters are divisors of L_i, respectively. The set Λ itself is given as

$$\Lambda = \left\{ (\boldsymbol{k}, \boldsymbol{l}) = (k_1, k_2, \ell_1, \ell_2) = (\alpha_1 u_1, \alpha_2 u_2, \beta_1 v_1, \beta_2 v_2) \,\middle|\, u_i \in \left\langle \tfrac{L_i}{\alpha_i} \right\rangle, v_i \in \left\langle \tfrac{L_i}{\beta_i} \right\rangle \right\},$$

i.e., it is a product group: $\Lambda = \Lambda_1 \times \Lambda_2$ with $\Lambda_i = \alpha_i \mathbb{Z}_{L_i} \times \beta_i \widehat{\mathbb{Z}_{L_i}}$.

Full separability may be violated in different ways. Assume that $\Lambda = \Lambda_1 \times \Lambda_2$ but with non-separable 2D lattices Λ_i. There are at least two natural choices, whose usefulness may depend on the concrete application. The first and probably more relevant choice is a lattice Λ_1 in position space and Λ_2, another lattice, in the wave-number domain. For the case of radial symmetric windows, g, one may choose a hexagonal packing in both the spatial and the wave-number domain.

Another flavor of separability comes in by choosing lattices within $\mathbb{C}^{L_1^2}$ and $\mathbb{C}^{L_2^2}$, respectively, describing the first and the second pair of phase space variables.

In passing, we note that there are also fully non-separable subgroups. They will not be discussed here, because it is not clear whether the increased level of technicality is worth the effort.

Separable Atoms on Fully Separable Lattices

In this section, we will show why the case of a 2D separable window $\boldsymbol{g} = g_1 \otimes g_2$ and a fully separable PF-lattice

$$\Lambda = \Lambda_1 \times \Lambda_2 = \alpha_1 \mathbb{Z}_{L_1} \times \beta_1 \widehat{\mathbb{Z}_{L_1}} \times \alpha_2 \mathbb{Z}_{L_2} \times \beta_2 \widehat{\mathbb{Z}_{L_2}}$$

allows for very efficient Gabor expansions at decent redundancy. It is crucial to observe that in this case, it is enough to find a dual 1D window γ_1 for the 1D window g_1 on the TF-lattice $\Lambda_1 \trianglelefteq \mathbb{Z}_{L_1} \times \widehat{\mathbb{Z}_{L_1}}$ and a dual 1D window γ_2 for the 1D window g_2 on the TF-lattice $\Lambda_2 \trianglelefteq \mathbb{Z}_{L_2} \times \widehat{\mathbb{Z}_{L_2}}$ in order to obtain a dual 2D window $\boldsymbol{\gamma}$ for $\boldsymbol{g}$ for the lattice Λ, simply as $\boldsymbol{\gamma} := \gamma_1 \otimes \gamma_2$. In short, the 2D Gabor frame on the product space $\mathbb{C}^{L_1} \otimes \mathbb{C}^{L_2}$ is obtained by combining via tensorization the Gabor frames for the signal spaces $\mathbb{C}^{L_1}$ and $\mathbb{C}^{L_2}$. The abstract result in the background can be summarized as follows:

Lemma 4. *If* $\{e_m\}_{m\in\langle N_1\rangle} \subseteq \mathbb{C}^{L_1}$ *and* $\{f_n\}_{n\in\langle N_2\rangle} \subseteq \mathbb{C}^{L_2}$ *are frames for* $\mathbb{C}^{L_1}$ *and* $\mathbb{C}^{L_2}$, *respectively, then the sequence* $\{e_m \otimes f_n\}_{(m,n)\in\langle N_1\rangle\times\langle N_2\rangle}$ *is a frame for* $\mathbb{C}^{L_1} \otimes \mathbb{C}^{L_2}$, *where* $(g \otimes h)(j,k) := g(j)\,h(k)$ *for* $g \in \mathbb{C}^{L_1}$ *and* $h \in \mathbb{C}^{L_2}$. *The joint redundancy is* $\frac{N_1N_2}{L_1L_2} \geq 1$.

As our image space is a tensor product, we define 2D Gabor windows $\boldsymbol{g} \in \mathbb{C}^{L_1\times L_2}$ by $\boldsymbol{g} = g_1 \otimes g_2$ for $g_i \in \mathbb{C}^{L_i}$. As we are looking at the case where $\Lambda = \Lambda_1 \times \Lambda_2$, we take two Gabor frames $\left\{g^{(i)}_{k_i,\ell_i}\right\}_{(k_i,\ell_i)\in\Lambda_i} := \{M_{\ell_i} T_{k_i} g_i\}_{(k_i,\ell_i)\in\Lambda_i} \subseteq \mathbb{C}^{L_i}$ with frame operators S_i and use the set of products $\left\{g^{(1)}_{k_1,\ell_1} \otimes g^{(2)}_{k_2,\ell_2}\right\}_{(\boldsymbol{k},\boldsymbol{l})\in\Lambda} \subseteq \mathbb{C}^{L_1} \otimes \mathbb{C}^{L_2}$ as frame for the image space with frame operator $S_1 \otimes S_2$.

In order to ensure the fact that this is a 2D Gabor family, one just has to verify that the translation by some element in a product group, applied to a tensor product, can be split into the action of each component to the corresponding factor. Finally, the exponential law implies that a similar splitting is valid for the modulation operators; in fact, plane waves are themselves tensor products of pure frequencies. We thus have altogether

$$M_{\ell_1} T_{k_1} g_1 \otimes M_{\ell_2} T_{k_2} g_2 = M_{(\ell_1,\ell_2)} T_{(k_1,k_2)} (g_1 \otimes g_2) \quad \forall (k_1,k_2),(\ell_1,\ell_2) \in \mathbb{Z}_{L_1} \times \mathbb{Z}_{L_2}$$

as building blocks for our 2D Gabor frame.

The canonical dual of $\boldsymbol{g}$ with respect to that frame is given as

$$\boldsymbol{\gamma}^\circ = S^{-1}\boldsymbol{g} = S_1^{-1} g_1 \otimes S_2^{-1} g_2 = \gamma_1^\circ \otimes \gamma_2^\circ.$$

The calculation of 1D dual windows for separable TF-lattices has been efficiently implemented in MATLAB available from the NuHAG webpage (http://www.univie.ac.at/nuhag-php/mmodule/resp. by Peter Søndergaard LTFAT (linked with the above page)).

Next, let us check out how we can efficiently obtain the Gabor coefficients of an image $\boldsymbol{f} \in \mathbb{C}^{L_1\times L_2}$ as given by (47). How does the Gabor matrix $\boldsymbol{C_g}$ look like if it is to be applied to an image $\boldsymbol{f} \in \mathbb{C}^{L_1\times L_2}$ stored as an $L_1 \times L_2$-matrix? For sure, $\boldsymbol{f}$ must be seen as a vector in $\mathbb{C}^{L_1L_2}$ and $\boldsymbol{C_g}$ as an $N_1N_2 \times L_1L_2$-matrix if the number of elements in the 2D frame is N_1N_2 and the coefficient vector is $c \in \mathbb{C}^{N_1N_2}$. In general, $\boldsymbol{f}$ cannot be assumed to be separable; thus, the only thing simplifying our computation is the structure

$$c_{\boldsymbol{k},\boldsymbol{l}} = \left\langle \boldsymbol{f}, M_{\ell_1} T_{k_1} g_1 \otimes M_{\ell_2} T_{k_2} g_2 \right\rangle_{\mathrm{F}}.$$

If we think of the 1D case with some $f \in \mathbb{C}^L$ and a general frame $\{g_j\}_{j\in\langle N\rangle} \subseteq \mathbb{C}^L$, the coefficients are obtained by

$$c = Cf = \left(\langle f, g_j\rangle\right)_{j\in\langle N\rangle} = (c_j)_{j\in\langle N\rangle},$$

and for Gabor frames, $c = (c_{k,\ell})_{(k,\ell)\in\Lambda}$ with $\Lambda \trianglelefteq \mathbb{Z}_L \times \widehat{\mathbb{Z}_L}$ is actually a coefficient matrix in $\mathbb{C}^{L\times L}$ with $|\Lambda| = N \leq L^2$ nonzero entries. But due to simply stacking the vectors $\{g_{k,\ell}\}_{(k,\ell)\in\Lambda} = \{g_j\}_{j\in\langle N\rangle} \subseteq \mathbb{C}^L$ in the coefficient matrix

$$C = \begin{pmatrix} g_0^* \\ \vdots \\ g_{N-1}^* \end{pmatrix} \in \mathbb{C}^{N\times L}, \tag{48}$$

one just gets a "flat" $c \in \mathbb{C}^N$. In our 2D case, the Gabor coefficient even consists of entries $c_{\boldsymbol{kl}} = c_{k_1,k_2,\ell_1,\ell_2}$. We also want to take the approach by using general frames $\{g_m\}_{m\in\langle N_1\rangle} \subseteq \mathbb{C}^{L_1}$ and $\{h_n\}_{n\in\langle N_2\rangle} \subseteq \mathbb{C}^{L_2}$ and look at the product frame $\{g_m \otimes h_n\}_{m,n}$ for $\mathbb{C}^{L_1} \otimes \mathbb{C}^{L_2}$. We also reduce the coefficient $c = (c(m,n))_{m,n} \in \mathbb{C}^{N_1N_2}$ to a vector of the form

$$\begin{aligned} c = \big(&c(0,0), c(0,1), \ldots, c(0,N_2-1), c(1,0), \ldots, c(1,N_2-1), \ldots \\ &\ldots, c(N_1-1,0), \ldots, c(N_1-1,N_2-1)\big)^{\mathrm{T}} \end{aligned}$$

such that we can try to find the corresponding coefficient matrix $C \in \mathbb{C}^{N_1N_2\times L_1L_2}$ that can be applied to $f \in \mathbb{C}^{L_1L_2}$, where

$$f = \big(f(0,0), \ldots, f(0,L_2-1), f(1,0), \ldots, f(L_1-1,L_2-1)\big)^{\mathrm{T}}. \tag{49}$$

Now, we can look at the (m,n)-th or, rather, (mN_2+n)-th entry of the coefficient:

$$\begin{aligned} (Cf)_{m,n} = c(m,n) &= \langle f, g_m \otimes h_n\rangle_{\mathbb{C}^{L_1L_2}} \\ &= \sum_{u=0}^{L_1-1}\sum_{v=0}^{L_2-1} f(u,v)\,\overline{(g_m\otimes h_n)(u,v)} \\ &= \sum_{u=0}^{L_1-1}\sum_{v=0}^{L_2-1} f(u,v)\,\overline{g_m(u)\,h_n(v)}. \end{aligned} \tag{50}$$

Since we are now able to split the indices u and v for the frame elements, we can consider the order in (49) and get

$$(Cf)_{m,n} = \big(\overline{g_m(0)}\,h_n^*\ \overline{g_m(1)}\,h_n^*\ \cdots\ \overline{g_m(L_1-1)}\,h_n^*\big)\,f = (C)_{m,n}f,$$

where $(C)_{m,n}$ is the (m,n)-th or (mN_2+n)-th line of C and contains L_1L_2 entries. The line vectors $\{h_n^*\}_{n\in\langle N_2\rangle}$ form the frame matrix $C_2 \in \mathbb{C}^{N_2\times L_2}$ like in (48). If we look at the range of N_2 lines $\{(m,0),\ldots,(m,N_2-1)\}$, we are able to express the corresponding segment of C as

$$(C)_{m;n\in\langle N_2\rangle} = \big(\overline{g_m(0)}\,C_2\ \overline{g_m(1)}\,C_2\ \cdots\ \overline{g_m(L_1-1)}\,C_2\big).$$

This shows that the frame matrix of the product frame is the Kronecker product of the partial frame operators $C_i \in \mathbb{C}^{N_i \times L_i}$, i = 1,2:

$$C = C_1 \otimes C_2 \in \mathbb{C}^{N_1 N_2 \times L_1 L_2}.$$

Nevertheless, we want to see whether we can compute $c = (C_1 \otimes C_2) f$ in a cheaper way by applying the frame matrices C_i without computing their Kronecker product. As images are not stored as vectors $f \in \mathbb{C}^{L_1 L_2}$ but rather as matrices $\boldsymbol{f} \in \mathbb{C}^{L_1 \times L_2}$ in numerical software like MATLAB or Octave, we could try to get the coefficient $\boldsymbol{c} = (c(m,n))_{m,n} \in \mathbb{C}^{N_1 \times N_2}$ more directly.

Proposition 1. *Given two frames $\{g_m\}_{m \in \langle N_1 \rangle} \subseteq \mathbb{C}^{L_1}$ and $\{h_n\}_{n \in \langle N_2 \rangle} \subseteq \mathbb{C}^{L_2}$ with frame matrices $C_i \in \mathbb{C}^{N_i \times L_i}$, then the frame coefficient $\boldsymbol{c} \in \mathbb{C}^{N_1 \times N_2}$ for the image $\boldsymbol{f} \in \mathbb{C}^{L_1 \times L_2}$ with respect to the product frame $\{g_m \otimes h_n\}_{(m,n)}$ is given by matrix multiplication as follows:*

$$\boldsymbol{c} = C_1 * \boldsymbol{f} * C_2^{\mathrm{T}} =$$

$$\begin{pmatrix} \overline{g_0(0)} & \cdots & \overline{g_0(L_1-1)} \\ \vdots & \vdots & \vdots \\ \overline{g_{N_1-1}(0)} & \cdots & \overline{g_{N_1-1}(L_1-1)} \end{pmatrix} \begin{pmatrix} f(0,0) & \cdots & f(0,L_2-1) \\ \vdots & \cdots & \vdots \\ f(L_1-1,0) & \cdots & f(L_1-1,L_2-1) \end{pmatrix} \begin{pmatrix} \overline{h_0(0)} & \cdots & \overline{h_{N_2-1}(0)} \\ \vdots & \cdots & \vdots \\ \overline{h_0(L_2-1)} & \cdots & \overline{h_{N_2-1}(L_2-1)} \end{pmatrix} \tag{51}$$

Note that similar thoughts reveal the fact that the 2D DFT of an image $\boldsymbol{f} \in \mathbb{C}^{L_1 \times L_2}$ can be obtained by the matrix multiplication

$$\mathcal{F}\boldsymbol{f} = F_{L_1} * \boldsymbol{f} * F_{L_2} \in \mathbb{C}^{L_1 \times L_2}, \tag{52}$$

where $F_{L_i} \in \mathbb{C}^{L_i \times L_i}$ are the (symmetric) Fourier matrices of order L_i.

If the synthesis operation is to be done by $f = C^* c$ for given $f \in \mathbb{C}^L$ and a frame $C \in \mathbb{C}^{N \times L}$, one solution is obtained by $c = (C^*)^\dagger f$ with a right inverse for C^* such that $\mathrm{I}_L = SS^{-1} = C^* C (C^* C)^{-1} = C^* (C^*)^\dagger$, making the pseudo-inverse of the synthesis operator the matching analysis operator. $C^*(C^*)^\dagger$ is the orthogonal projection onto the range of the desired synthesis operator. One notices that due to $(C^*)^\dagger = (C^\dagger)^*$, we already have $\mathrm{I}_L = (C^\dagger C)^* = C^\dagger C$, the orthogonal projection onto the range of ran $C^\dagger$. Thus, the role of the operators can be interchanged, meaning that $C^\dagger$ is the matching synthesis operator for the analysis operator C.

If we again interpret signals $\boldsymbol{f} \in \mathbb{C}^{L_1} \otimes \mathbb{C}^{L_2}$ as $f \in \mathbb{C}^{L_1 L_2}$ and take a product frame $\{g_m \otimes h_n\}_{m,n}$ with analysis operator $C_1 \otimes C_2$, we get $\mathrm{I}_{L_1 L_2} = C^\dagger (C_1 \otimes C_2)$ and $\mathrm{I}_{L_1 L_2} = \mathrm{I}_{L_1} \otimes \mathrm{I}_{L_2} = \left(C_1^\dagger C_1\right) \otimes \left(C_2^\dagger C_2\right)$, yielding that the matching synthesis operator is $C^\dagger = C_1^\dagger \otimes C_2^\dagger$. Due to Proposition 1, we can thus reconstruct $\boldsymbol{f} \in \mathbb{C}^{L_1 \times L_2}$ by

$$\boldsymbol{f} = \left(C_1^\dagger C_1\right)\boldsymbol{f}\left(C_2^\dagger C_2\right)^{\mathrm{T}} = C_1^\dagger \boldsymbol{c}\left(C_2^\dagger\right)^{\mathrm{T}} \tag{53}$$

because $\boldsymbol{c} = C_1 \boldsymbol{f} C_2^{\mathrm{T}}$ is in the range of the corresponding analysis operator.

These results were derived for products of general frames and therefore also hold for products of Gabor frames. Given two Gabor frames $\{M_{\ell_i} T_{k_i} g_i\}$ on subgroups $\Lambda_i \trianglelefteq \mathbb{Z}_{L_i} \times \widehat{\mathbb{Z}_{L_i}}$ and with analysis operators C_{g_i}, we get their synthesis operators by $C_{g_i}^\dagger = C_{\gamma_i^\circ}^*$ with $\gamma_i^\circ := S_{g_i}^{-1} g_i$. The product of those two frames is the Gabor frame $\{\boldsymbol{M_l T_k g}\}_{(\boldsymbol{k},\boldsymbol{l})\in\Lambda_1\times\Lambda_2}$ consisting of PF-shifts of the window $\boldsymbol{g} = g_1 \otimes g_2 \in \mathbb{C}^{L_1\times L_2}$ on the lattice $\Lambda = \Lambda_1 \times \Lambda_2$. The dual window to $\boldsymbol{g}$ is given by $\boldsymbol{\gamma}^\circ := \gamma_1^\circ \otimes \gamma_2^\circ$. Due to (51) and (53), the 2D Gabor analysis operation for the image $\boldsymbol{f} \in \mathbb{C}^{L_1\times L_2}$ is obtained by

$$\boldsymbol{c} = C_{g_1} \boldsymbol{f} C_{g_2}^{\mathrm{T}} \tag{54}$$

and a possible reconstructing synthesis operation by

$$\boldsymbol{f} = C_{\gamma_1^\circ}^* \boldsymbol{c}\left(C_{\gamma_2^\circ}^*\right)^{\mathrm{T}} = \overline{C_{\gamma_1^\circ}}^{\mathrm{T}} \boldsymbol{c}\,\overline{C_{\gamma_2^\circ}}, \tag{55}$$

yielding that it is enough to obtain the two duals γ_i°. Figure 3 shows the construction and look of the separable dual 2D window of a 2D Gaussian window on a fully separable PF-lattice.

Efficient Gabor Expansion by Sampled STFT

In the case of a separable 2D atom and a fully separable PF-lattice, we can make use of any fast 1D STFT implementation (cf. the NuHAG software page or the LTFAT by Peter Søndergaard) to obtain the Gabor analysis coefficient $\boldsymbol{c} = C_{g_1} \boldsymbol{f} C_{g_2}^{\mathrm{T}}$ and the Gabor reconstruction $\boldsymbol{f} = C_{\gamma_1^\circ}^* \boldsymbol{c}\left(C_{\gamma_2^\circ}^*\right)^{\mathrm{T}}$ for a given image $\boldsymbol{f} \in \mathbb{C}^{L_1\times L_2}$. These matrix multiplications from the left and right could still be rather expensive, so one can obtain the set of Gabor coefficients $\boldsymbol{c}$ by calculating a finite number of sampled 1D STFTs, with the sampling points determined by shift parameters α_1, α_2 and modulation parameters β_1, β_2.

If we remember the 1D case, the Gabor frame C_g for $\mathbb{C}^L$ by a window $g \in \mathbb{C}^L$ involves a separable lattice $\Lambda = \alpha\mathbb{Z}_L \times \beta\mathbb{Z}_L$ with $|\Lambda| = N = \frac{L^2}{\alpha\beta}$, and for arbitrary $f \in \mathbb{C}^L$, we have

$$(C_g f)_{k,\ell} = c_{k,\ell} = \langle f, M_{\beta\ell} T_{\alpha k} g\rangle_{\mathbb{C}^L} = \sum_{u=0}^{L-1} f(u)\,\overline{M_{\beta\ell} T_{\alpha k} g(u)} = \mathcal{V}_g f(\alpha k, \beta\ell)$$

for $k \in \langle \frac{L}{\alpha} \rangle$ and $\ell \in \left\langle \frac{L}{\beta} \right\rangle$, which can be viewed as a vector of length N if the frame is seen as a matrix $C_g \in \mathbb{C}^{N \times L}$. In the 2D case, if we consider $\boldsymbol{f} = (f_0, \ldots, f_{L_2-1})$ with $f_j := (f(0,j), \ldots, f(L_1-1,j))^{\mathrm{T}}$, then $b_j = C_{g_1} f_j$ acts as the Gabor analysis operation for all $f_j \in \mathbb{C}^{L_1}$ with coefficients $b_j \in \mathbb{C}^{N_1}$ for all $j \in \langle L_2 \rangle$. The operation $\boldsymbol{b} = C_{g_1} \boldsymbol{f}$ collects these in a matrix $\boldsymbol{b} = (b_0, \ldots, b_{L_2-1})$. If we express its k-th line as a line vector $q_k^{\mathrm{T}} := (\boldsymbol{b})_k = (b_0(k), \ldots, b_{L_2-1}(k))$, we get

$$C_{g_1} \boldsymbol{f} = \boldsymbol{b} = \boldsymbol{q}^{\mathrm{T}} = \begin{pmatrix} q_0^{\mathrm{T}} \\ \vdots \\ q_{N_1-1}^{\mathrm{T}} \end{pmatrix} \in \mathbb{C}^{N_1 \times L_2}.$$

The complete 2D Gabor analysis operation is thus $\boldsymbol{c} = \boldsymbol{q}^{\mathrm{T}} C_{g_2}^{\mathrm{T}} = (C_{g_2} \boldsymbol{q})^{\mathrm{T}}$, and this is just the Gabor analysis operation of the vectors $q_k \in \mathbb{C}^{L_2}$ for $k \in \langle N_1 \rangle$ with respect to the Gabor frame C_{g_2}.

All in all, the 2D Gabor analysis operation in the twofold-separable case can be obtained by first computing L_2 1D STFT-operations of output length N_1 using the parameters α_1, β_1 followed by N_1 1D STFT-operations of output length N_2 using the parameters α_2, β_2.

As the reconstruction (2D Gabor expansion) is just a multiplication of the dual Gabor matrices $C_{\gamma_i}^*$ from the left and right of $\boldsymbol{c}$, this task can be seen as a sequence of 1D Gabor expansions and can thus be obtained by a sequence of inverse 1D STFT-operations as well. There are again two ways: The first one is to do N_1 inverse operations with output length L_2 using the parameters α_2, β_2 followed by N_2 operations with output length L_1 using α_1, β_1. The second way exchanges L_i and N_i correspondingly.

Visualizing a Sampled STFT of an Image

So far, we have visualized the full STFT of an image as a large block image, where either each block fully represents the frequency domain and the position of the blocks the position domain, or vice versa. As such an image would become rather huge, we prefer to visualize only a sampled STFT instead. In the case of a separable atom, this can be realized by obtaining the discrete 2D Gabor transform by (54), where the two involved matrices C_{g_i} consider a special order of their Gabor frame elements $M_{\ell_i} T_{k_i} g_i$.

For a Gabor frame $\{M_\ell T_k g\}_{(k,\ell) \in \Lambda} \subseteq \mathbb{C}^L$ given by a 1D window $g \in \mathbb{C}^L$ on a separable lattice $\Lambda = \alpha \mathbb{Z}_L \times \beta \mathbb{Z}_L$ with $N = |\Lambda| = \frac{L^2}{\alpha\beta}$ elements, we say that the Gabor frame elements are ordered by *modulation priority* if the frame matrix $C_g \in \mathbb{C}^{N \times L}$ is of the form

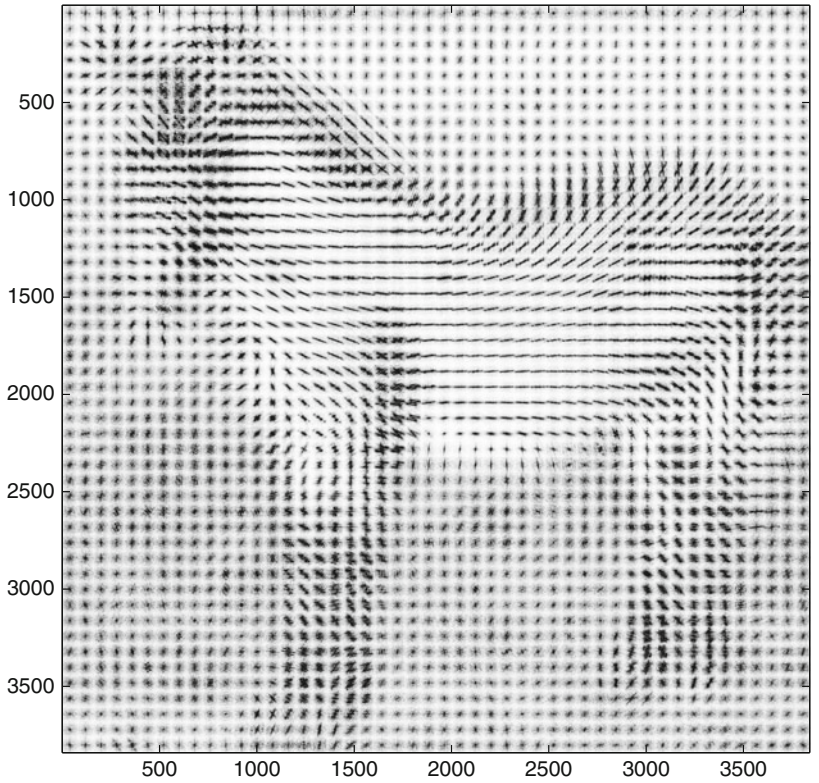


Fig. 4 Discrete 2D Gabor transform of a zebra, modulation priority. The picture shows the absolute values of $c = C_g f C_g^{\mathrm{T}}$, where g is the 1D Gaussian of length 480 and C_g is the Gabor matrix for the lattice $\Lambda = 10\,\mathbb{Z}_{480} \times 6\,\mathbb{Z}_{480}$, whose entries were ordered with modulation priority

$$C_g = \begin{pmatrix} M_0 T_0 g^* \\ M_\beta T_0 g^* \\ \vdots \\ M_{L/\beta-1} T_0 g^* \\ M_0 T_1 g^* \\ \vdots \\ M_{L/\beta-1} T_1 g^* \\ \vdots \\ M_{L/\beta-1} T_{L/\alpha-1} g^* \end{pmatrix}$$

We call it ordered by *translation priority* if it is of the form

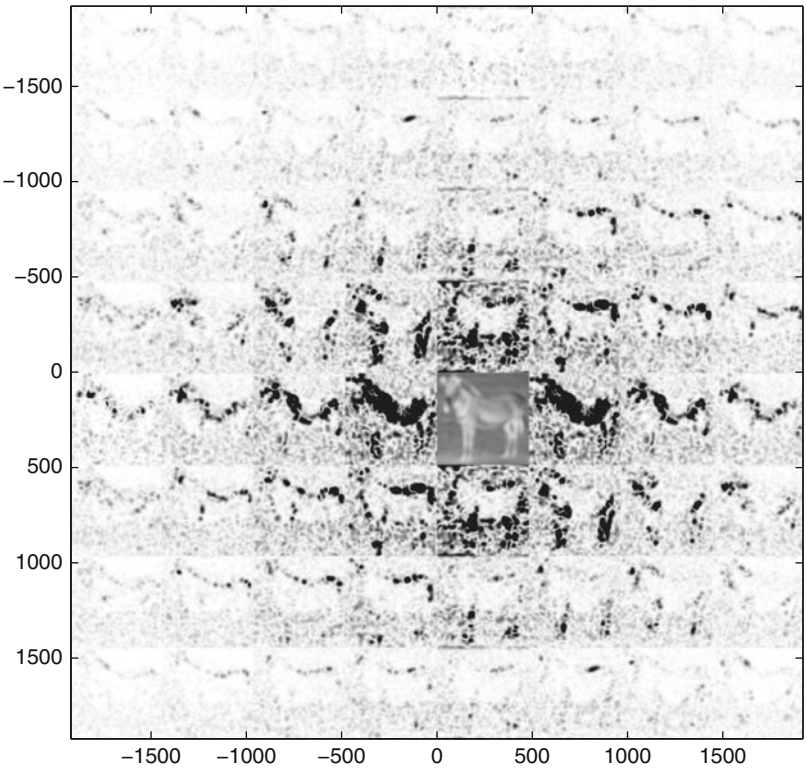


Fig. 5 Discrete 2D Gabor transform of a zebra, translation priority. The picture shows the absolute values of $\widetilde{c} = \widetilde{C}_g \boldsymbol{f} \widetilde{C}_g^{\mathrm{T}}$, where g is the 1D Gaussian of length 480 and $\widetilde{C}_g$ is the Gabor matrix for the lattice $\Lambda = \mathbb{Z}_{480} \times 60\,\mathbb{Z}_{480}$, whose entries were ordered with translation priority. The Gaussian blurred image in the middle has been scaled into the colormap individually

$$\tilde{C}_g = \begin{pmatrix} M_0 T_0 g^* \\ M_0 T_\alpha g^* \\ \vdots \\ M_0 T_{L/\alpha - 1} g^* \\ M_1 T_0 g^* \\ \vdots \\ M_1 T_{L/\alpha - 1} g^* \\ \vdots \\ M_{L/\beta - 1} T_{L/\alpha - 1} g^* \end{pmatrix}$$

Obviously, $\tilde{C}_g = P C_g$ for a suitable permutation matrix $P \in \mathbb{C}^{N \times N}$.

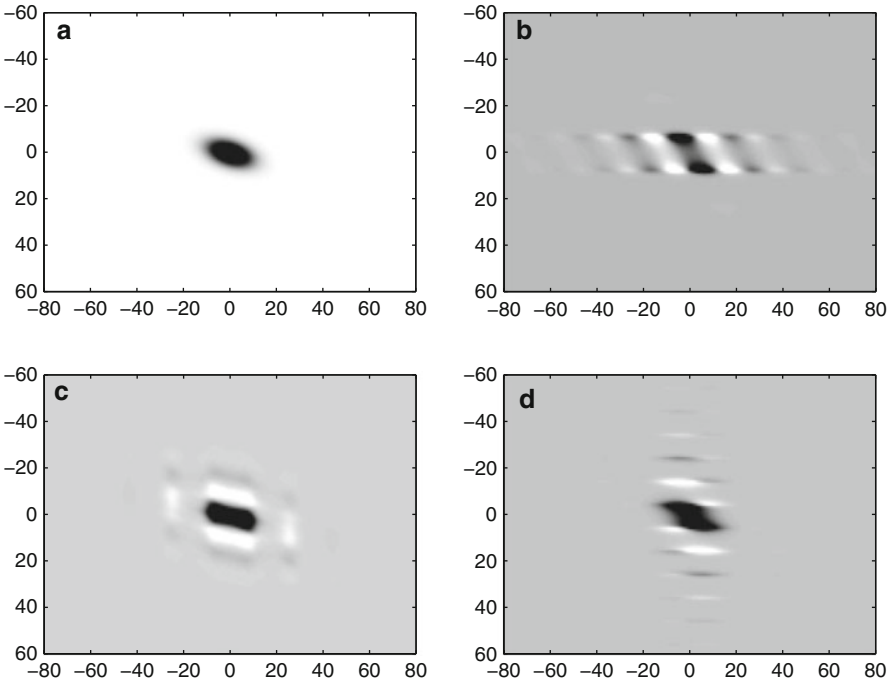


Fig. 6 A non-separable window and some duals on fully separable lattices. The lattices Λ_i are that of Fig. 3. The lattices Λ_i' exchange α_i with β_i. The last lattice has vertical redundancy 1 and horizontal redundancy 6.4. (**a**) Non-separable window g. (**b**) Dual γ° on $\Lambda_1 \times \Lambda_2$. (**c**) Dual on $\Lambda_1' \times \Lambda_2'$ with exchange parameters (**d**) Dual at vertically critical redundancy

If we take an image $\boldsymbol{f} \in \mathbb{C}^{L_1 \times L_2}$ and two Gabor frames $\{M_{\ell_i} T_{k_i} g_i\}$, $(k_i, \ell_i) \in \Lambda_i$, on separable lattices $\Lambda_i = \alpha_i \mathbb{Z}_{L_i} \times \beta_i \mathbb{Z}_{L_i}$, we can take their product Gabor frame for $\mathbb{C}^{L_1 \times L_2}$ and obtain the mentioned two possibilities for an STFT block image by either considering the frame matrices C_{g_i} or $\tilde{C}_{g_i}$. The matrices C_{g_i} are ordered by modulation priority, and if $\boldsymbol{c} = C_{g_1} \boldsymbol{f} C_{g_2}^{\mathrm{T}}$, then $\boldsymbol{c}$ consists of $\frac{L_1}{\beta_1} \times \frac{L_2}{\beta_2}$-blocks

$$X_{k_1,k_2} := \left(\langle \boldsymbol{f}, M_{(\ell_1,\ell_2)} T_{(k_1,k_2)} \boldsymbol{g} \rangle \right)_{\ell_1,\ell_2}$$

such that

$$\boldsymbol{c} = \begin{pmatrix} X_{0,0} & \cdots & X_{0,L_2/\alpha_2-1} \\ \vdots & \cdots & \vdots \\ X_{L_1/\alpha_1-1,0} & \cdots & X_{L_1/\alpha_1-1,L_2/\alpha_2-1} \end{pmatrix}.$$

The blocks X_{k_1,k_2} equal the part $\left(\mathcal{V}_{\boldsymbol{g}} \boldsymbol{f}(k_1, k_2, \ell_1, \ell_2) \right)_{\ell_1,\ell_2}$ of the sampled STFT and thus contain the whole (sampled) set of frequency shifts for a certain position shift

of the window $\boldsymbol{g} = g_1 \otimes g_2$. The (sampled) frequency domain is therefore spanned in each of the blocks X_{k_1,k_2}, and their positions in $\boldsymbol{c}$ span the (sampled) position domain. Each X_{k_1,k_2} could be seen as a sampled "Fourier image" of the discrete Fourier transform $\widehat{\boldsymbol{f}} \cdot T_{(k_1,k_2)}\bar{\boldsymbol{g}}$.

In the other case, where we have $\tilde{\boldsymbol{c}} = \tilde{C}_{g_1} \boldsymbol{f} \tilde{C}_{g_2}^{\mathrm{T}}$, the Gabor coefficient consists of $\frac{L_1}{\alpha_1} \times \frac{L_2}{\alpha_2}$-blocks

$$Y_{\ell_1,\ell_2} := \left(\langle \boldsymbol{f}, M_{(\ell_1,\ell_2)} T_{(k_1,k_2)} \boldsymbol{g} \rangle\right)_{k_1,k_2}$$

such that

$$\tilde{\boldsymbol{c}} = \begin{pmatrix} Y_{0,0} & \cdots & Y_{0,L_2/\beta_2-1} \\ \vdots & \cdots & \cdots \\ Y_{L_1/\beta_1-1,0} & \cdots & Y_{L_1/\beta_1-1,L_2/\beta_2-1} \end{pmatrix}.$$

Here, the blocks Y_{ℓ_1,ℓ_2} equal the part $\left(\mathcal{V}_{\boldsymbol{g}} \boldsymbol{f}(k_1, k_2, \ell_1, \ell_2)\right)_{k_1,k_2}$ of the sampled STFT and contain the corresponding set of position shifts for a certain frequency shift of $\boldsymbol{g}$. The position domain is spanned in each of the blocks Y_{ℓ_1,ℓ_2}, and their positions in $\tilde{\boldsymbol{c}}$ span the frequency domain.

Figures 4 and 5 show examples for both cases using the zebra test image. As it is a square image, we can take $g_1 = g_2$ and thus $C_{g_1} = C_{g_2}$. The first figure composes the Gabor transform coefficient matrix as blocks of Fourier images. Clearly, the overall image reflects the shape of the zebra. The "pixels" of that image contain "Fourier jets" that are orthogonal to the edges at the corresponding position in the original zebra image. Thus, the "jets" are oriented horizontally where, e.g., the body of the animal shows vertical line patterns. The second figure shows blocks of zebra images that have been convolved with modulated Gaussians. The absolute values show the peaks as black spots within the respective image blocks.

Non-separable Atoms on Fully Separable Lattices

Non-separable windows are those that can only be defined considering the complete image domain $\mathbb{Z}_{L_1} \times \mathbb{Z}_{L_2}$, and not $\mathbb{Z}_{L_1}$ and $\mathbb{Z}_{L_2}$ separately. These cannot be described as a tensor product $g_1 \otimes g_2$ with $g_i \in \mathbb{C}^{L_i}$ anymore, but only generally as $\boldsymbol{g} \in \mathbb{C}^{L_1 \times L_2}$. With this case, we lose the ability to consider two (1D) frames independently for each dimension, and we cannot apply two frame matrices independently to an image. It appears that we have to stick to the known factorizations of Gabor matrices on (fully) separable lattices with parameters α_i, β_i, and we thus cannot make use of the equidistantly sampled 1D STFT. However, under certain conditions, this case can be completely referred to a 1D case, as we will see below.

Figure 6 indicates an important thing about the redundancy. Sure, a redundancy of $\frac{|\Lambda|}{L_1 L_2} \geq 1$ is only a necessary condition, but it seems to be important to consider

the redundancy in each dimension. The involved window is a 2D Gaussian window $\boldsymbol{g} \in \mathbb{C}^{120\times 160}$, stretched vertically by $\frac{4}{3}$, shrunken horizontally by $\frac{3}{4}$, then rotated (counter-clockwise) by $\frac{3}{8}\pi$. Figure 6d shows its dual on a fully separable 4D PF-lattice with overall redundancy 6.4. It was computed in the work of P. Prinz which makes use of the Gabor matrix factorizations [66]. But although the redundancy value gives the impression to be safe, it hides the fact that the involved lattice is actually $\Lambda = 10\,\mathbb{Z}_{120} \times 12\,\mathbb{Z}_{120} \times 5\,\mathbb{Z}_{160} \times 5\,\mathbb{Z}_{160}$, yielding the redundancy as $\frac{120}{10\cdot 12} \cdot \frac{160}{5\cdot 5} = 1 \cdot 6.4$. This shows that the vertical redundancy is critical, and the dual has a bad localization in the vertical dimension. It is therefore necessary to make sure that the redundancy is reasonably distributed among the dimensions. In this sense, fully separable 4D lattices can always be considered as a product of two 2D TF-lattices with independent redundancies, no matter what structure the 2D window possesses.

7 Historical Notes and Hint to the Literature

Nonorthogonal expansions as proposed by D. Gabor in his seminal work [45] of 1946 were ignored for a long time by the mathematical community. The question to which extent the claims made by D. Gabor could be realized in the context of generalized functions, was carefully analyzed by A.J.E.M. Janssen in 1981 [53]. Around the same time, M. Bastiaans explored the connections between Gabor theory and optics [3–7]. In the critically sampled case, he suggested to use the biorthogonal function γ in order to calculate Gabor coefficients. The connection to the biorthogonality relations for dual Gabor windows was pointed out in two papers in 1995 [26, 54] and brought to the multidimensional case in [39, 40, 69].

Two early papers in the field, authored by J. Daugman and Y.Y. Zeevi and his coauthors, established a connection between a 2D version of Gabor analysis and early vision [27, 46, 64, 83, 84], Various subsequent papers emphasized that a Gabor family is not an orthogonal system and that, therefore, computation of coefficients has to be computationally expensive. We know by now that while linear independence is indeed lost, the rich covariance structure of the Gabor problems actually leads to efficient algorithms.

The mathematical theory of Gabor expansions was promoted in various directions in the last two decades. Although a lot of Gabor analysis is naturally valid in the context of general locally compact Abelian groups, a substantial body of references only covers the standard case, for 1D signals and separable lattices.

Of course, the theory underlying image processing is formally covered by the theory of Gabor analysis over finite Abelian groups as described in [40]. Some basic facts in the general LCA context are given in [50], and some further results generalize to this setting, applying standard facts from abstract harmonic analysis [44].

Multidimensional, non-separable lattices are discussed in [41], and [38] deals with situations where the isomorphism of 2D groups with certain 1D groups helps to use 1D Gabor code to calculate 2D dual Gabor windows.

Numerical methods for Gabor representations have been discussed since the first and pioneering papers (see, e.g., [2,46,84]). There are also hints on how to perform parallel versions of the Gabor transform [30]. A partial comparison of algorithms is in [67] and in the toolbox of P. Søndergaard. It can be expected to provide further implementations and more details concerning numerical issues in the near future.

One of the most natural applications (based on the interpretation of Gabor coefficients) are *space-variant filters*. Given the Gabor transform, one can multiply them with a $0/1$ function over the coefficient domain, passing through, e.g., higher frequencies within regions of interest, whereas otherwise, only low frequencies are stored, thus representing foveated images (with somewhat blurred parts outside the region of interest).

Since different textures in different regions of an image might also be detected using Gabor coefficients, natural applications are texture segmentation (see e.g., [31,77]), image restoration [19,82], and image fusion [68]. The extraction of directional features in images has been considered recently in [48]. Other contributions to texture analysis are found in [49]. Other applications are pattern recognition [76], face identification as described in [70], and face detection [52].

Some of the material presented in this paper can be found in an extended form in the master thesis of the last named author [63].

Cross-References

▸ Compressive Sensing
▸ Energy Minimization Methods
▸ Starlet Transform in Astronomical Data Processing

References

1. Ali, S.T., Antoine, J.-P., Murenzi, R., Vandergheynst, P.: Two-Dimensional Wavelets and Their Relatives. Cambridge University Press, Cambridge (2004)
2. Assaleh, K., Zeevi, Y., Gertner, I.: On the realization of Zak-Gabor representation of images. In: Proceedings of SPIE: Visual Communications and Image Processing'91, Boston, vol. 1606, pp. 532–552. SPIE (1991)
3. Bastiaans, M.J.: Gabor's expansion of a signal into Gaussian elementary signals. Proc. IEEE **68**(4), 538–539 (1980)
4. Bastiaans, M.J.: A sampling theorem for the complex spectrogram and Gabor's expansion of a signal in Gaussian elementary signals. Opt. Eng. **20**(4), 594–598 (1981)
5. Bastiaans, M.J.: On the sliding-window representation in digital signal processing. IEEE Trans. Acoust. Speech Signal Process. **33**(4), 868–873 (1985)
6. Bastiaans, M.J.: Application of the Wigner distribution function to partially coherent light. J. Opt. Soc. Am. **3**(8), 1227–1238 (1986)
7. Bastiaans, M.J.: Gabor's signal expansion in optics. In: Feichtinger, H.G., Strohmer, T. (eds.) Gabor Analysis and Algorithms: Theory and Applications. Applied and Numerical Harmonic Analysis, pp. 427–451. Birkhäuser, Boston (1998)
8. Bastiaans, M.J., van Leest, A.J.: From the rectangular to the quincunx Gabor lattice via fractional Fourier transformation. IEEE Signal Proc. Lett. **5**(8), 203–205 (1998)

9. Bastiaans, M.J., van Leest, A.J.: Product forms in Gabor analysis for a quincunx-type sampling geometry. In: Veen, J. (ed.) Proceedings of the CSSP-98, ProRISC/IEEE Workshop on Circuits, Systems and Signal Processing, Mierlo, 26–17 Nov 1998. STW, Technology Foundation, Utrecht, pp. 23–26 (1998)
10. Battle, G.: Heisenberg proof of the Balian-Low theorem. Lett. Math. Phys. **15**(2), 175–177 (1988)
11. Ben Arie, J., Rao, K.R.: Nonorthogonal signal representation by Gaussians and Gabor functions. IEEE Trans Circuits-II **42**(6), 402–413 (1995)
12. Ben Arie, J., Wang, Z.: Gabor kernels for affine-invariant object recognition. In: Feichtinger, H.G., Strohmer, T. (eds.) Gabor Analysis and Algorithms: Theory and Applications. Birkhauser, Boston (1998)
13. Bölcskei, H., Feichtinger, H.G., Gröchenig, K., Hlawatsch, F.: Discrete-time multi-window Wilson expansions: pseudo frames, filter banks, and lapped transforms. In: Proceedings of the IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis, Paris, pp. 525–528 (1996)
14. Bölcskei, H., Gröchenig, K., Hlawatsch, F., Feichtinger, H.G.: Oversampled Wilson expansions. IEEE Signal Proc. Lett. **4**(4), 106–108 (1997)
15. Bölcskei, H., Janssen, A.J.E.M.: Gabor frames, unimodularity, and window decay. J. Fourier Anal. Appl. **6**(3), 255–276 (2000)
16. Christensen, O.: An Introduction to Frames and Riesz Bases. Applied and Numerical Harmonic Analysis. Birkhäuser, Boston (2003)
17. Christensen, O.: Frames and Bases: An Introductory Course. Applied and Numerical Harmonic Analysis. Birkhäuser, Basel (2008)
18. Coifman, R.R., Matviyenko, G., Meyer, Y.: Modulated Malvar-Wilson bases. Appl. Comput. Harmonic Anal. **4**(1), 58–61 (1997)
19. Cristobal, G., Navarro, R.: Blind and adaptive image restoration in the framework of a multiscale Gabor representation. In: Proceedings of the IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis, Philadelphia, pp. 306–309 (1994)
20. Cristobal, G., Navarro, R.: Space and frequency variant image enhancement based on a Gabor representation. Pattern Recognit. Lett. **15**(3), 273–277 (1994)
21. Cvetkovic, Z., Vetterli, M.: Oversampled filter banks. IEEE Trans. Signal Process. **46**(5), 1245–1255 (1998)
22. Daubechies, I.: Time-frequency localization operators: a geometric phase space approach. IEEE Trans. Inf. Theory **34**(4), 605–612 (1988)
23. Daubechies, I.: The wavelet transform, time-frequency localization and signal analysis. IEEE Trans. Inf. Theory **36**(5), 961–1005 (1990)
24. Daubechies, I., Grossmann, A., Meyer, Y.: Painless nonorthogonal expansions. J. Math. Phys. **27**(5), 1271–1283 (1986)
25. Daubechies, I., Jaffard, S., Journé, J.L.: A simple Wilson orthonormal basis with exponential decay. SIAM J. Math. Anal. **22**, 554–573 (1991)
26. Daubechies, I., Landau, H.J., Landau, Z.: Gabor time-frequency lattices and the Wexler-Raz identity. J. Fourier Anal. Appl. **1**(4), 437–478 (1995)
27. Daugman, J.G.: Complete discrete 2-D Gabor transforms by neural networks for image analysis and compression. IEEE Trans. Acoust. Speech Signal Process. **36**(7), 1169–1179 (1988)
28. Dubiner, Z., Porat, M.: Position-Variant Filtering in the Position Frequency Space: Performance Analysis and Filter Design, pp. 1–34 (1997)
29. Dufaux, F., Ebrahimi, T., Geurtz, A., Kunt, M.: Coding of digital TV by motion-compensated Gabor decomposition. In: Tescher, A.G. (ed.) Proceedings of SPIE: Applications of Digital Image Processing XIV, Image Compression, San Diego, 22 July 1991, vol. 1567, pp. 362–379. SPIE (1991)
30. Dufaux, F., Ebrahimi, T., Kunt, M.: Massively parallel implementation for real-time Gabor decomposition. In: Tzou, K.-H., Koga, T. (eds.) Proceedings of SPIE: Proceedings of Visual

Communications and Image Processing'91: Image Processing, Boston. Volume 1606 of VLSI Implementation and Hardware Architectures, pp. 851–864. SPIE (1991)
31. Dunn, D., Higgins, W.E.: Optimal Gabor filters for texture segmentation. IEEE Trans. Image Process. **4**(7), 947–964 (1995)
32. Ebrahimi, T., Kunt, M.: Image compression by Gabor expansion. Opt. Eng. **30**(7), 873–880 (1991)
33. Ebrahimi, T., Reed, T.R., Kunt, M.: Video coding using a pyramidal Gabor expansion. In: Proceedings of SPIE: Visual Communications and Image Processing'90, Lausanne, vol. 1360, pp. 489–502. SPIE (1990)
34. Feichtinger, H.G.: Modulation spaces: looking back and ahead. Sampl. Theory Signal Image Process. **5**(2), 109–140 (2006)
35. Feichtinger, H.G., Gröchenig, K.: Banach spaces related to integrable group representations and their atomic decompositions, I. J. Funct. Anal. **86**, 307–340 (1989)
36. Feichtinger, H.G., Gröchenig, K.: Theory and practice of irregular sampling. In: Benedetto, J., Frazier, M. (eds.) Wavelets: Mathematics and Applications. Studies in Advanced Mathematics, pp. 305–363. CRC, Boca Raton (1994)
37. Feichtinger, H.G., Gröchenig, K., Walnut, D.F.: Wilson bases and modulation spaces. Math. Nachr. **155**, 7–17 (1992)
38. Feichtinger, H.G., Kaiblinger, N.: 2D-Gabor analysis based on 1D algorithms. In: Proceedings of the OEAGM-97, Hallstatt (1997)
39. Feichtinger, H.G., Kozek, W.: Quantization of TF lattice-invariant operators on elementary LCA groups. In: Feichtinger, H.G., Strohmer, T. (eds.) Gabor Analysis and Algorithms. Theory and Applications. Applied and Numerical Harmonic Analysis, pp. 233–266, 452–488. Birkhäuser, Boston (1998)
40. Feichtinger, H.G., Kozek, W., Luef, F.: Gabor analysis over finite Abelian groups. Appl. Comput. Harmonic Anal. **26**, 230–248 (2009)
41. Feichtinger, H.G., Kozek, W., Prinz, P., Strohmer, T.: On multidimensional non-separable Gabor expansions. In: Proceedings of SPIE: Wavelet Applications in Signal and Image Processing IV, Denver (1996)
42. Feichtinger, H.G., Luef, F., Werther, T.: A guided tour from linear algebra to the foundations of Gabor analysis. In: Gabor and Wavelet Frames. Volume 10 of Lecture Notes Series, Institute for Mathematical Sciences, National University of Singapore, pp. 1–49. World Scientific, Hackensack (2007)
43. Feichtinger, H.G., Strohmer, T., Christensen, O.: A group theoretical approach to Gabor analysis. Opt. Eng. **34**, 1697–1704 (1995)
44. Folland, G.B.: Harmonic Analysis in Phase Space. Princeton University Press, Princeton (1989)
45. Gabor, D.: Theory Commun. J. IEE **93**(26), 429–457 (1946)
46. Gertner, I., Zeevi, Y.Y.: Image representation with position-frequency localization. In: International Conference on Acoustics, Speech, and Signal Processing (ICASSP-91), Toronto, vol. 4, pp. 2353–2356 (1991)
47. Golub, G., van Loan, C.F.: Matrix Computations, 3rd edn. Johns Hopkins University Press, Baltimore (1996)
48. Grafakos, L., Sansing, C.: Gabor frames and directional time frequency analysis. Appl. Comput. Harmonic Anal. **25**(1), 47–67 (2008)
49. Grigorescu, S., Petkov, N., Kruizinga, P.: Comparison of texture features based on Gabor filters. IEEE Trans. Image Process. **11**(10), 1160–1167 (2002)
50. Gröchenig, K.: Aspects of Gabor analysis on locally compact Abelian groups. In: Feichtinger, H.G., Strohmer, T. (eds.) Gabor Analysis and Algorithms: Theory and Applications, pp. 211–231. Birkhäuser, Boston (1998)
51. Gröchenig, K.: Foundations of Time-Frequency Analysis. Applied and Numerical Harmonic Analysis. Birkhäuser, Boston (2001)

52. Hoffmann, U., Naruniec, J., Yazdani, A., Ebrahimi, T.: Face detection using discrete Gabor jets and a probabilistic model of colored image patches. In: Filipe, J., Obaidat, M.S. (eds.) E-Business and Telecommunications (ICETE 2008), Porto, 26–29 July 2008. Revised Selected Papers, Volume 48 of Communications in Computer and Information Science, pp. 331–344 (2008)
53. Janssen, A.J.E.M.: Gabor representation of generalized functions. J. Math. Anal. Appl. **83**, 377–394 (1981)
54. Janssen, A.J.E.M.: Duality and biorthogonality for Weyl-Heisenberg frames. J. Fourier Anal. Appl. **1**(4), 403–436 (1995)
55. Janssen, A.J.E.M.: From continuous to discrete Weyl-Heisenberg frames through sampling. J. Fourier Anal. Appl. **3**(5), 583–596 (1997)
56. Kutyniok, G., Strohmer, T.: Wilson bases for general time-frequency lattices. SIAM J. Math. Anal. **37**(3), 685–711 (electronic) (2005)
57. Lee, T.S.: Image representation using 2D Gabor wavelets. IEEE Trans. Pattern Anal. Mach. Intell. **18**(10), 959–971 (1996)
58. Li, S.: Discrete multi-Gabor expansions. IEEE Trans. Inf. Theory **45**(6), 1954–1967 (1999)
59. Lu, Y., Morris, J.: Fast computation of Gabor functions. IEEE Signal Process. Lett. **3**(3), 75–78 (1996)
60. Malvar, H.S.: Lapped transforms for efficient transform/subband coding. IEEE Trans. Acoust. Speech Signal Process. **38**(6), 969–978 (1990)
61. Navarro, R., Portilla, J., Tabernero, A.: Duality between oveatization and multiscale local spectrum estimation. In: Rogowitz, B.E., Pappas, T.N. (eds.) Proceedings of SPIE: Human Vision and Electronic Imaging III, San Jose, 26 Jan 1998, vol. 3299, pp. 306–317. SPIE, Bellingham (1998)
62. Nestares, O., Navarro, R., Portilla, J., Tabernero, A.: Efficient spatial-domain implementation of a multiscale image representation based on Gabor functions. J. Electron. Imaging **7**(1), 166–173 (1998)
63. Paukner, S.: Foundations of Gabor analysis for image processing. Master's thesis, University of Vienna (2007)
64. Porat, M., Zeevi, Y.Y.: The generalized Gabor scheme of image representation in biological and machine vision. IEEE Trans. Pattern Anal. Mach. Intell. **10**(4), 452–468 (1988)
65. Porat, M., Zeevi, Y.Y.: Gram-Gabor approach to optimal image representation. In: Kunt, M. (ed.) Proceedings of SPIE: Visual Communications and Image Processing'90: Fifth in a Series, Lausanne, vol. 1360, pp. 1474–1478. SPIE (1990)
66. Prinz, P.: Calculating the dual Gabor window for general sampling sets. IEEE Trans. Signal Process. **44**(8), 2078–2082 (1996)
67. Redding, N., Newsam, G.: Efficient calculation of finite Gabor transforms. IEEE Trans. Signal Process. **44**(2), 190–200 (1996)
68. Redondo, R., Sroubek, F., Fischer, S., Cristobal, G.: Multifocus image fusion using the log-Gabor transform and a Multisize Windows technique. Inf. Fusion **10**(2), 163–171 (2009)
69. Ron, A., Shen, Z.: Weyl-Heisenberg frames and Riesz bases in L2(Rd). Duke Math. J. **89**(2), 237–282 (1997)
70. Shen, L., Bai, L., Fairhurst, M.: Gabor wavelets and general discriminant analysis for face identification and verification. Image Vis. Comput. **25**(5), 553–563 (2007)
71. Søndergaard, P.L.: Finite discrete Gabor analysis. PhD thesis, Technical University of Denmark (2007)
72. Strohmer, T.: Numerical algorithms for discrete Gabor expansions. In: Feichtinger, H.G., Strohmer, T. (eds.) Gabor Analysis and Algorithms: Theory and Applications, pp. 267–294. Birkhäuser, Boston (1997)
73. Subbanna, N.K., Zeevi, Y.Y.: Image representation using noncanonical discrete multi-window Gabor frames. In: IET International Conference on Visual Information Engineering (VIE 2006), Bangalore, pp. 482–487 (2006)
74. Urieli, S., Porat, M., Cohen, N.: Optimal reconstruction of images from localized phase. IEEE Trans. Image Process. **7**(6), 838–853 (1998)

75. van Leest, A.J., Bastiaans, M.J.: Gabor's signal expansion and the Gabor transform on a non-separable time-frequency lattice. J. Frankl. Inst. **337**(4), 291–301 (2000)
76. Vargas, A., Campos, J., Navarro, R.: An application of the Gabor multiscale decomposition of an image to pattern recognition. In: Proceedings of SPIE: Visual Communications and Image Processing'96, Orlando, vol. 2730, pp. 622–625. SPIE (1996)
77. Weldon, T., Higgins, W., Dunn, D.: Efficient Gabor filter design for texture segmentation. Pattern Recognit. **29**(12), 2005–2015 (1996)
78. Werner, D.: Funktionalanalysis. (Functional Analysis) 2., Überarb. Au. Springer, Berlin (1997)
79. Wojdyllo, P.: Modified Wilson orthonormal bases. Sampl. Theory Signal Image Process. **6**(2), 223–235 (2007)
80. Wojdyllo, P.: Characterization of Wilson systems for general lattices. Int. J. Wavelets Multiresolut. Inf. Process. **6**(2), 305–314 (2008)
81. Wojtaszczyk, P.: Stability and instance optimality for Gaussian measurements in compressed sensing. Found. Comput. Math. **10**, 1–13 (2010)
82. Yang, J., Liu, L., Jiang, T., Fan, Y.: A modified Gabor filter design method for fingerprint image enhancement. Pattern Recognit. Lett. **24**(12), 1805–1817 (2003)
83. Zeevi, Y.Y.: Multiwindow Gabor-type representations and signal representation by partial information. In: Byrnes, J.S. (ed.) Twentieth Century Harmonic Analysis – A Celebration Proceedings of the NATO Advanced Study Institute, II, Ciocco, 2–15 July 2000. Volume 33 of NATO Science Series II, Mathematics, Physics, and Chemistry, pp. 173–199. Kluwer, Dordrecht (2001)
84. Zeevi, Y.Y., Zibulski, M., Porat, M.: Multi-window Gabor schemes in signal and image representations. In: Feichtinger, H.G., Strohmer, T. (eds.) Gabor Analysis and Algorithms: Theory and Applications. Applied and Numerical Harmonic Analysis, pp. 381–407. Birkhäuser, Boston (1998)
85. Zibulski, M., Zeevi, Y.Y.: Matrix algebra approach to Gabortype image representation. In: Haskell, B.G., Hang, H.-M. (eds.) Proceedings of SPIE: Visual Communications and Image Processing'93, Wavelet, Cambridge, 08 Nov 1993, vol. 2094, pp. 1010–1020. SPIE (1993)
86. Zibulski, M., Zeevi, Y.Y.: Frame analysis of the discrete Gaborscheme. IEEE Trans. Signal Process. **42**(4), 942–945 (1994)
87. Zibulski, M., Zeevi, Y.Y.: Analysis of multiwindow Gabor-type schemes by frame methods. Appl. Comput. Harmonic Anal. **4**(2), 188–221 (1997)
88. Zibulski, M., Zeevi, Y.Y.: Discrete multiwindow Gabor-type transforms. IEEE Trans. Signal Process. **45**(6), 1428–1442 (1997)
89. Zibulski, M., Zeevi, Y.Y.: The generalized Gabor scheme and its application in signal and image representation. In: Signal and Image Representation in Combined Spaces. Volume 7 of Wavelet Analysis and Its Applications, pp. 121–164. Academic, San Diego (1998)

Shape Spaces

Alain Trouvé and Laurent Younes

Contents

1 Introduction 1760
2 Background 1760
3 Mathematical Modeling and Analysis 1762
Some Notation 1762
A Riemannian Manifold of Deformable Landmarks 1763
Hamiltonian Point of View 1772
Spaces of Plane Curves 1786
Extension to More General Shape Spaces 1800
Applications to Statistics on Shape Spaces 1802
4 Numerical Methods and Case Examples 1803
Landmark Matching via Shooting 1804
Landmark Matching via Path Optimization 1807
Computing Geodesics Between Curves 1808
Inexact Matching and Optimal Control Formulation 1810
5 Conclusion 1813
Cross-References 1813
References 1814

Abstract

This chapter describes a selection of models that have been used to build Riemannian spaces of shapes. It starts with a discussion of the finite-dimensional

A. Trouvé (✉)
Centre de Mathématiques et Leurs Applications, Ecole Normale Superieure Cachan, Cachan Cédex, France
e-mail: trouve@cmla.ens-cachan.fr

L. Younes
Department of Applied Mathematics and Statistics, The Johns Hopkins University, Baltimore, MD, USA
e-mail: laurent.younes@jhu.edu

© Springer Science+Business Media New York 2015

O. Scherzer (ed.), *Handbook of Mathematical Methods in Imaging*,
DOI 10.1007/978-1-4939-0790-8_55

space of point sets (or landmarks) and then provides an introduction to the more challenging issue of building spaces of shapes represented as plane curves. A special attention is devoted to constructions involving quotient spaces, since they are involved in the definition of shape spaces via the action of groups of diffeomorphisms and in the process of identifying shapes that can be related by a Euclidean transformation. The resulting structure is first described via the geometric concept of a Riemannian submersion and then reinterpreted in a Hamiltonian and optimal control framework, via momentum maps. These developments are followed by the description of algorithms and illustrated by numerical experiments.

1 Introduction

The analysis of shapes as mathematical objects has constituted a significant area of interest in the past few decades motivated by the development of image acquisition methods and segmentation algorithms, in which shapes could be extracted as isolated objects. Shape analysis is a framework, in which a given shape is considered as a single (typically infinite dimensional) variable, requiring the development of new techniques for their representation and statistical interpretation. This framework has found applications in several fields, including object recognition in computer vision and computational anatomy.

The example in Fig. 1 can help in framing the kind of problems that are being addressed and serve as a motivation. These shapes are fairly easily recognizable for the human eye. They do however exhibit large variations, and a description in simple terms of how they vary and of how they can be compared is a much harder task. It is clear that a naive representation, like a list of points, cannot be used directly, because the discretized curves may have different numbers of points, and no correspondence is available between them. Coming up with quantitative and reliable descriptors that can be, for example, analyzed in a rigorous statistical study is, however, of main importance for the many applications, and the goal of this chapter is to provide a framework in which such a task can be performed in a reliable well-posed way.

2 Background

During the past decades, several essential contributions have been made, using rigorous mathematical concepts and methods, to address this problem and others of similar nature. This collection of efforts has progressively defined a new discipline that can be called *mathematical shape theory*.

Probably, the first milestone in the development of the theory is Kendall's construction of a space of shapes, defined as a quotient of the space of disjoint points in $\mathbb{R}^d$ by the action of translation, rotation, and scaling [40]. Kendall's theory has been the starting point of a huge literature [15, 41, 64] and allowed for new approaches for studying datasets in which the group of similitudes was a nuisance

Fig. 1 Examples of shapes (Taken from the MPEG-7 shape database)

factor (for such data as human skulls, prehistoric jewelry, etc.). One can argue that, as a candidate for a shape space, Kendall's model suffers from two main limitations. First, it relies on the representation of a shape by a finite number of labeled points, or *landmarks*. These landmarks need to have been identified on each shape, and shapes with different numbers of landmarks belong to different spaces. From a practical point of view, landmarks are most of the time manually selected, the indexation of large datasets being time consuming and prone to user-dependent errors. The second limitation is that the metric on shapes is obtained by quotienting out the standard Euclidean metric on point sets, using a standard "Riemannian submersion" process that we will discuss later in this chapter. The Euclidean metric ignores a desirable property of shape comparison, which states that shapes that are smooth deformations of one another should be considered more similar than those for which the points in correspondence are randomly displaced, even if the total point displacement is the same.

This important issue, related to smoothness, was partially addressed by another important contribution to the theory, which is Bookstein's use of the thin plate splines originally developed by Duchon and Meinguet [10, 16, 51]. Splines interpolate between *landmark displacements* to obtain a smooth, dense, displacement field (or vector field). It can be addressed with the generic point of view of *reproducing kernel Hilbert spaces* [7, 77], which will also be reviewed later in this chapter.

This work had a tremendous influence on shape analysis based on landmarks, in particular for medical studies. It suffers, however, from two major drawbacks. The first one is that the interpolated displacement can be ambiguous, with several points moved to the same position. This is an important limitation, since inferring unobserved correspondences is one of the objectives of this method. The second drawback, in relation with the subject of this chapter, is that the linear construction associated to splines fails to provide a metric structure on the nonlinear space of shapes. The spline deformation energy provides in fact a first-order approximation of a nonconstant Riemannian metric on point sets, which provides an interesting version of a manifold of landmarks, as introduced in [11, 36, 72].

After point sets, plane curves are certainly the shape representation in which most significant advances have been observed over the last few years. Several important metrics have been discussed in publications like [37, 42, 43, 52, 79–81]. They have been cataloged, among many other metrics, in a quasiencyclopedic effort by D. Mumford and P. Michor [55]. We will return to some of these metrics in section "Spaces of Plane Curves."

Grenander's theory of deformable templates [27] is another seminal work for shape spaces. In a nutshell, Grenander's basic idea, which can be traced back to D'Arcy Thomson's work on biological shapes in the beginning of last century [65], is to introduce suitable group actions as generative engines for visual object models, with the natural use of the group of diffeomorphisms for shapes. While the first developments in this context use linear approximations of diffeomorphisms [2, 28, 29], a first computational breakthrough in the nonlinear estimation of diffeomorphisms was provided in [12] with the introduction of flows associated to ordinary differential equations. This idea was further developed in a fully metric approach of diffeomorphisms and shape spaces, in a framework that was introduced in [17, 66, 67] and further developed in [8, 11, 35, 36, 57, 58]. The approach also led to important developments in medical imaging, notably via the establishment of a new discipline, called computational anatomy, dedicated to the study of datasets of anatomical shapes [30, 60, 61, 78].

3 Mathematical Modeling and Analysis

Some Notation

The following notation will be used in this chapter. The Euclidean norm of vectors $a \in \mathbb{R}^d$ will be denoted using single bars and the dot product between a and b as $a \cdot b$ or explicitly as $a^T b$, where a^T is the transpose of a. So

$$|a|^2 = a \cdot a = a^T a$$

for $a \in \mathbb{R}^d$.

Other norms (either Riemannian metrics or norms on infinite-dimensional spaces) will be denoted with double bars, generally with a subscript indicating the corresponding space, or relevant point in the manifold. We will use angles for the corresponding inner product, with the same index, so that, for a Hilbert space V, the notation for the inner product between V and w in V will be $\langle v\,,\, w\rangle_V$ with

$$\|v\|_V^2 = \langle v\,,\, v\rangle_V.$$

When f is a function that depends on a variable t, its derivative with respect to t computed at some point t_0 will be denoted either $\partial_t f(t_0)$ or $\dot{f}_t(t_0)$, depending on which form gives the most readable formula. Primes are never used to denote derivative, that is, f' is not the derivative of f, but just another function. The differential at x of a function of several variables F is denoted $DF(x)$. If F is scalar valued, its gradient is denoted $\nabla F(x)$. The divergence of a vector field $v : \mathbb{R}^d \to \mathbb{R}^d$ is denoted $\nabla \cdot v$.

If M is a differential manifold, the tangent space to M at $x \in M$ will be denoted $T_x M$ and its cotangent space (dual of the former) $T_x^* M$. The tangent bundle (disjoint union of the tangent spaces) is denoted TM and the cotangent bundle T^*M.

When μ is a linear form on a vector space V (i.e., a scalar-valued linear transformation), the natural pairing between μ and $v \in V$ will be denoted $(\mu \mid v)$, that is,

$$(\mu \mid v) = \mu(v).$$

A Riemannian Manifold of Deformable Landmarks

Interpolating Splines and RKHSs

Let us start with some preliminary facts on Hilbert spaces of functions or vector fields and their relation with interpolating splines. A Hilbert space is a possibly infinite-dimensional vector space equipped with an inner product which induces a complete topology. Letting V be such a space, with norm and inner product, respectively, denoted $\|\cdot\|_V$ and $\langle \cdot\,,\,\cdot \rangle_V$, a linear form on V is a continuous linear transformation $\mu : V \mapsto \mathbb{R}$. The set of such transformations is called the dual space of V and denoted V^*. An element μ in V^* being continuous by definition, there exists a constant C such that

$$\forall v \in V, \mu(v) \leq C \|v\|_V.$$

The smaller number C for which this assertion is true is called the operator norm of μ and denoted $\|\mu\|_{V^*}$.

Instead of $\mu(v)$ like above, the notation $(\mu \mid v)$ will be used to represent the result of μ applied to V. The Riesz representation theorem implies that V^* is in one-to-one correspondence with V, so that for any μ in V^*, there exists a unique element $v = K_V \mu \in V$ such that, for any $w \in V$,

$$(\mu \mid w) = \langle K_V \mu\,,\, w \rangle_V;$$

K_V and its inverse $L_V = K_V^{-1}$ are called the duality operators of V. They provide an isometric identification between V and V^*, with, in particular, $\|\mu\|^2_{V^*} = (\mu \mid K_V \mu) = \|K_V \mu\|^2_V$.

Of particular interest is the case when V is a space of vector fields in d dimensions, that is, of functions $v : \mathbb{R}^d \to \mathbb{R}^d$ (or from $\Omega \to \mathbb{R}^d$ where Ω is an open subset of $\mathbb{R}^d$), and when the norm in V is such that the evaluation functionals $a \otimes \delta_x$ belong to V^* for any $a, x \in \mathbb{R}^d$, where

$$(a \otimes \delta_x \mid v) = a^T v(x), v \in V. \tag{1}$$

In this case, the vector field $K_V(a \otimes \delta_x)$ is well defined and linear in a. One can define the matrix-valued function $(y, x) \mapsto \tilde{K}_V(y, x)$ by

$$\tilde{K}_V(y, x)a = (K_V(a \otimes \delta_x))(y);$$

$\tilde{K}_V$ is the kernel of the space V. In the following, we will write $K_V(x, y)$ instead of $\tilde{K}_V(x, y)$, with the customary abuse of notation of identifying the kernel and the operator that it defines.

One can easily deduce from its definition that K_V satisfies the reproducing property

$$\forall a, b \in \mathbb{R}^d, \langle K_V(\cdot, x)a\ ,\ K_V(\cdot, y)b\rangle_V = a^T K_V(x, y)b,$$

which also implies the symmetry property $K_V(x, y) = K_V(y, x)^T$. Unless otherwise specified, it will always be assumed that V is a space of vector fields that vanish at infinity, which implies the same property for the kernel (one variable tending to infinity and the other remaining fixed).

A space V as considered above is called a reproducing kernel Hilbert space (RKHS) of vector fields. Fixing such a space, one can consider the *spline interpolation problem*, which is to find $v \in V$ with minimal norm such that $v(x_i) = c_i$, where $x_1, \ldots, x_N$ are points in $\mathbb{R}^d$ and $c_1, \ldots, c_N$ are d-dimensional vectors. It is quite easy to prove that the solution takes the form

$$v(y) = \sum_{i=1}^{N} K_V(y, x_i)\alpha_i, \tag{2}$$

where $\alpha_1, \ldots, \alpha_N$ are identified by solving the dN-dimensional system

$$\sum_{i=1}^{N} K_V(x_j, x_i)\alpha_i = c_j, \text{ for } j = 1, \ldots, N. \tag{3}$$

Let $S_V(\mathbf{x})$ (where $\mathbf{x} = (x_1, \ldots, x_N)$) denote the dN by dN block matrix

$$S_V(\mathbf{x}) = (K_V(x_i, x_j))_{i,j=1,\ldots,N}.$$

Stacking $c_1, \ldots, c_N$ and $\alpha_1, \ldots, \alpha_N$ in dN-dimensional column vectors $\mathbf{c}$ and $\boldsymbol{\alpha}$, one can show that, for the optimal V,

$$\|v\|_V^2 = \boldsymbol{\alpha}^T S_V(\mathbf{x})\boldsymbol{\alpha} = \mathbf{c}^T S(\mathbf{x})_V^{-1}\mathbf{c}, \tag{4}$$

each term representing this spline deformation energy for the considered interpolation problem.

How one uses this interpolation method now depends on how one interprets the vector field V. One possibility is to consider it as a *displacement field*, in the sense that a particle at position x in space is moved to position $x + v(x)$, therefore involving the space transformation $\varphi^v := \mathrm{id} + v$. In this view, the interpolation problem can be rephrased as finding the smoothest (in the V-norm sense) full space interpolation of given landmark displacements. The deformation energy in

(4) can then be interpreted as some kind of "elastic" energy that evaluates the total stress involved in the transformation φ^v. This (with some variants, including allowing for some no-cost affine, or polynomial, transformations) is the framework of interpolation based on thin plates, or radial basis functions, as introduced in [3, 4, 9, 10, 18], for example. As discussed in the introduction, this approach does not lead to a nice mathematical notion of a shape space of landmarks; moreover, in the presence of large displacements, the interpolated transformation φ^v may fail to be one to one and therefore to provide a well-defined dense correspondence.

The other way to interpret V is as a velocity field, so that $v(x)$ is the speed of a particle at x at a given time. The interpolation problem is then to obtain a smooth velocity field given the speeds $c_1, \ldots, c_N$ of particles $x_1, \ldots, x_N$. This point of view has the double advantage of providing a diffeomorphic displacement when the velocity field is integrated over time and allowing for the interpretation of the deformation energy as a kinetic energy, directly related to a Riemannian metric on the space of landmarks.

Riemannian Structure

Let Lmk_N denote the submanifold of $\mathbb{R}^{dN}$ consisting of all ordered collections of N distinct points in $\mathbb{R}^d$:

$$Lmk_N = \{\mathbf{x} = (x_1, \ldots, x_N) \in (\mathbb{R}^d)^N, x_i \neq x_j \text{ if } i \neq j\}.$$

The tangent space to Lmk_N at $\mathbf{x}$ can be identified to the space of all families of d-dimensional vectors $\mathbf{c} = (c_1, \ldots, c_N)$, and one defines (with the same notation as in the previous section) the Riemannian metric on Lmk_N

$$\|\mathbf{c}\|^2_{\mathbf{x}} = \mathbf{c}^T S_V(\mathbf{x})^{-1}\mathbf{c}.$$

As already pointed out, $\|\mathbf{c}\|^2_{\mathbf{x}}$ is the minimum of $\|v\|^2_V$ among all V in V such that $v(x_i) = c_i, i = 1, \ldots, N$. This minimum is attained at

$$v^{\mathbf{c}}(\cdot) = \sum_{i=1}^{N} K(\cdot, x_j)\alpha_j$$

with $\boldsymbol{\alpha} = S_V(\mathbf{x})^{-1}\mathbf{c}$.

Now, given any differentiable curve $t \mapsto \mathbf{x}(t)$ in Lmk_N, one can build an optimal time-dependent velocity field

$$v(t, \cdot) = v^{\mathbf{c}(t)}(\cdot)$$

with $\mathbf{c} = \partial_t \mathbf{x}$. One can then define the flow associated to this time-dependent velocity, namely, the time-dependent diffeomorphism φ^v, such that $\varphi^v(0, x) = x$ and

$$\partial_t \varphi^v(t, x) = v(t, \varphi^v(t, x))$$

which is, by construction, such that $\varphi^v(t, x_i(0)) = x_i(t)$ for $i = 1, \ldots, N$. So, this construction provides a diffeomorphic extrapolation of any curve in Lmk_N, which is optimal in the sense that its velocity has minimal V norms, given the induced constraints. The metric that has been defined on Lmk_N is the projection of the V norm via the infinitesimal action of velocity fields on Lmk_N, which is defined by

$$v \cdot (x_1, \ldots, x_N) = (v(x_1), \ldots, v(x_N)).$$

This concept will be extensively discussed later on in this chapter.

Geodesic Equation

Geodesics on Lmk_N are curves that locally minimize the energy, that is, they are curves $t \mapsto \mathbf{x}(t)$ such that, for any t, there exists $h > 0$ such that

$$\int_{t-h}^{t+h} \|\dot{\mathbf{x}}_u(u)\|^2_{\mathbf{x}(u)} du$$

is minimal over all possible curves in Lmk_N that connect $\mathbf{x}(t-h)$ and $\mathbf{x}(t+h)$. The geodesic, or Riemannian, distance between $\mathbf{x}_0$ and $\mathbf{x}_1$ is defined as the minimizer of the square root of the geodesic energy

$$\int_0^1 \|\dot{\mathbf{x}}_u\|^2_{\mathbf{x}(u)} du$$

over all curves in Lmk_N that connect $\mathbf{x}_0$ and $\mathbf{x}_1$.

Geodesics are characterized by a second-order equation, called the geodesic equation. If one denotes $G_V(\mathbf{x}) = S_V(\mathbf{x})^{-1}$, with coefficients $g_{(k,i),(l,j)}$ for $k, l = 1, \ldots, N$ and $i, j = 1, \ldots, d$, the classical expression of this equation is

$$\ddot{x}_{k,i} + \sum_{l,l'=1'}^{N} \sum_{j,j'=1}^{d} \Gamma^{(k,i)}_{(l,j),(l',j')} \dot{x}_{l,j} \dot{x}_{l',j'} = 0,$$

where $\Gamma^{(k,i)}_{(l,j),(l',j')}$ are the Christoffel symbols, given by

$$\Gamma^{(k,i)}_{(l,j),(l',j')} = \frac{1}{2}\Big(\partial_{x_{l',j'}} g_{(k,i),(l,j)} + \partial_{x_{l,j}} g_{(k,i),(l',j')} - \partial_{x_{k,i}} g_{(l,j),(l',j')}\Big).$$

In these formulae, the two indices that describe the coordinates in Lmk_N, $x_{k,i}$ were made explicit, representing the ith coordinate of the kth landmark. Solutions of this equation are unique as soon as $\mathbf{x}(0)$ and $\dot{\mathbf{x}}(0)$ are specified.

Equations put in this form become rapidly intractable when the number of landmarks becomes large. The inversion of the matrix $S_V(\mathbf{x})$ or even simply its storage can be computationally impossible when N gets larger than a few thousands.

It is much more efficient, and analytically simpler as well, to use the *Hamiltonian form* of the geodesic equation, which is (see [38])

$$\begin{cases} \partial_t \mathbf{x} = S_V(\mathbf{x})\boldsymbol{\alpha} \\ \partial_t \boldsymbol{\alpha} = -\dfrac{1}{2}\partial_{\mathbf{x}}(\boldsymbol{\alpha}^T S_V(\mathbf{x})\boldsymbol{\alpha}) \end{cases} \tag{5}$$

This equation will be justified in section "General Principles," in which the optimality conditions for geodesics will be retrieved as a particular case of general problems in calculus of variations and optimal control. Its solution is uniquely defined as soon as $\mathbf{x}(0)$ and $\boldsymbol{\alpha}(0)$ are specified. The time-dependent collection of vectors $t \mapsto \boldsymbol{\alpha}(t)$ is called the *momentum* of the motion. It is related to the velocity $\mathbf{c}(t) = \dot{\mathbf{x}}(t)$ by the identity $\mathbf{c} = S_V(\mathbf{x})\boldsymbol{\alpha}$.

Introducing K_V and letting $K_V^{ij}(x, y)$ denote the i, j entry of $K_V(x, y)$, this geodesic equation can be rewritten in the following even more explicit form:

$$\begin{cases} \partial_t x_k = \displaystyle\sum_{l=1}^{N} K_V(x_k, x_l)\alpha_l, \quad k = 1, \ldots, N, \\ \partial_t \alpha_k = -\displaystyle\sum_{l=1}^{N}\sum_{i,j=1}^{d} \nabla_1 K_V^{ij}(x_k, x_l)\alpha_{k,i}\alpha_{l,j}, \quad k = 1, \ldots, N, \end{cases} \tag{6}$$

where $\nabla_1 K_V^{ij}$ denotes the gradient of the i, j entry of K_V with respect to its first variable.

The geodesic equation defines the Riemannian exponential map as follows. Fix $\mathbf{x}_0 \in Lmk_N$. The exponential map at $\mathbf{x}_0$ is the transformation $\mathbf{c} \mapsto \exp_{\mathbf{x}_0}(\mathbf{c})$ defined over all tangent vectors $\mathbf{c}$ to Lmk_N at $\mathbf{x}_0$ (which are identified to all families of d-dimensional vectors, $\mathbf{c} = (c_1, \ldots, c_N)$), such that $\exp_{\mathbf{x}_0}(\mathbf{c})$ is the solution at time $t = 1$ of the geodesic equation initialized at $\mathbf{x}(0) = \mathbf{x}_0$ and $\dot{\mathbf{x}}(0) = \mathbf{c}$. Alternatively, one can define the exponential chart in Hamiltonian form that will also be called the *momentum representation* in Lmk_N by the transformation

$$\boldsymbol{\alpha}_0 \mapsto \exp^{\flat}_{\mathbf{x}_0}(\boldsymbol{\alpha}_0),$$

where $\exp^{\flat}_{\mathbf{x}_0}(\boldsymbol{\alpha}_0)$ is the solution at time 1 of system (6) initialized at $(\mathbf{x}_0, \boldsymbol{\alpha}_0)$.

For the metric that is considered here, one can prove that the exponential map at $\mathbf{x}_0$ (resp. the momentum representation) is defined for any vector $\mathbf{c}$ (resp. $\boldsymbol{\alpha}_0$); this also implies that they both are onto, so that any landmark configuration $\mathbf{y}$ can be written as $\mathbf{y} = \exp^{\flat}_{\mathbf{x}_0}(\boldsymbol{\alpha}_0)$ for some $\boldsymbol{\alpha}_0 \in (\mathbb{R}^d)^N$. The representation is not one to one, because geodesics may intersect, but it is so if restricted to a small-enough neighborhood of 0. More precisely, there exists an open subset $U \subset T_{\mathbf{x}_0} Lmk_N$ over which $\exp_{\mathbf{x}_0}$ is a diffeomorphism. This provides the so-called *exponential chart* at $\mathbf{x}$ on the manifold.

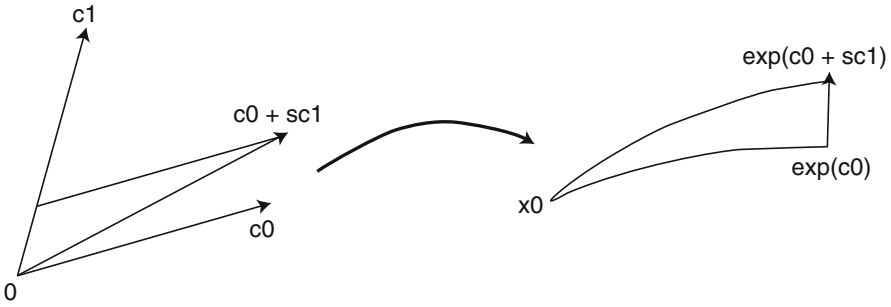


Fig. 2 Metric distortion for the exponential chart

Metric Distortion and Curvature

Exponential charts are often used for data analysis on a manifold, because they provide, in a neighborhood of a reference point $\mathbf{x}_0$, a vector-space representation which has no radial metric distortion, in the sense that, in the chart, the geodesic distance between $\mathbf{x}_0$ and $\exp_{\mathbf{x}_0}(\mathbf{c})$ is equal to $\|\mathbf{c}\|_{\mathbf{x}_0}$. The representation does distort the metric in the other directions. One way to measure this is by comparing (see Fig. 2), for given $\mathbf{c}_0$ and $\mathbf{c}_1$ with $\|\mathbf{c}_0\|_{\mathbf{x}_0} = \|\mathbf{c}_1\|_{\mathbf{x}_0} = 1$, the points $\exp_{\mathbf{x}_0}(t\mathbf{c}_0)$ and $\exp_{\mathbf{x}_0}(t(\mathbf{c}_0 + s\mathbf{c}_1))$. Let $F(t,s)$ denote the last term (so that the first one is $F(t,0)$). One can write

$$\mathrm{dist}(F(t,s), F(t,0)) = s\|\partial_s F(t,0)\|_{F(t,0)} + o(s).$$

Without metric distortion, this distance would be given by $st\|\mathbf{c}_1\|_{\mathbf{x}_0} = st$. However, it turns out that [14]

$$\|\partial_s F(t,0)\|_{F(t,0)} = t - \rho_{\mathbf{x}}(\mathbf{c}_0, \mathbf{c}_1)\frac{t^3}{6} + o(t^3),$$

where $\rho_{\mathbf{x}}(\mathbf{c}_0, \mathbf{c}_1)$ is the *sectional curvature* of the plane generated by $\mathbf{c}_0$ and $\mathbf{c}_1$ in $T_{\mathbf{x}_0}Lmk_N$. So, this sectional curvature directly measures (among many other things) the first order of metric distortion in the manifold and is therefore an important indication of this distortion of the exponential charts.

The usual explicit formula for the computation of the curvature involves the second derivatives of the metric tensor matrix $G_V(\mathbf{x})$, which, as we have seen, is intractable for large values of N. In a recent work, Micheli [53] introduced an interesting new formula for the computation of the curvature in terms of the inverse tensor, $S_V(\mathbf{x}_0)$.

Invariance

The previous landmark space ignored the important facts that two shapes are usually considered as identical when one can be deduced from the other by a Euclidean transformation, which is a combination of a rotation and a translation

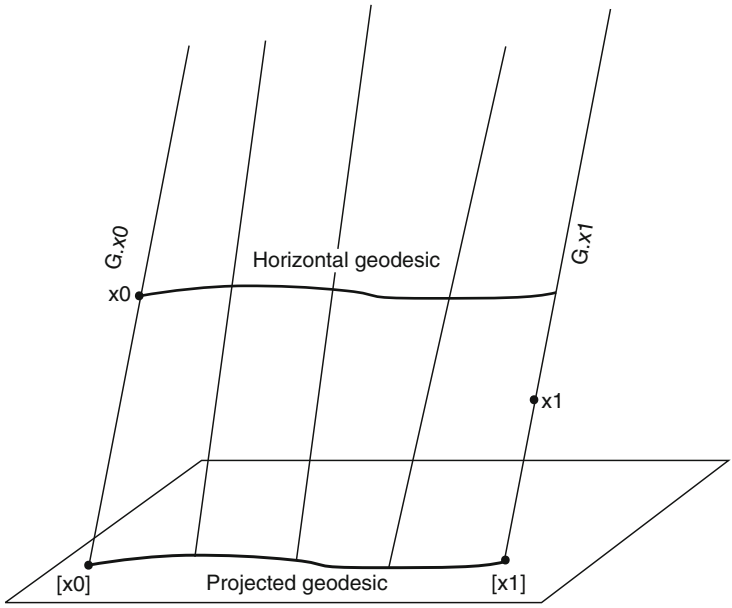


Fig. 3 Riemannian submersion (geodesics in the quotient space)

(scale invariance is another important aspect that will not be discussed in this section). To take this into account, we need to "mod out" these transformations, that is, to consider the quotient space of Lmk_N by the Euclidean group.

One can pass from the metric discussed in the previous section to a metric on the quotient space via the mechanism of Riemannian submersion (Fig. 3). The scheme is relatively simple, and we describe it and set up notation in a generic framework before taking the special case of the landmark manifold. So, let Q be a Riemannian manifold and $\pi : Q \to M$ be a submersion, that is, a smooth surjection from Q to another manifold M such that its differential $D\pi$ has full rank everywhere. This implies that, for $m \in M$, the set $\pi^{-1}(m)$ is a submanifold of Q, called the fiber at m. If $q \in Q$ and $m = \pi(q)$, the tangent space T_qQ can be decomposed into the direct sum of the tangent space to $\pi^{-1}(m)$ and the space perpendicular to it. We will refer to the former as the space of vertical vectors at q, and denote it $\mathcal{V}_q$, and to the latter as the space of horizontal vectors, denoted $\mathcal{H}_q$. We therefore have

$$T_qQ = \mathcal{V}_q \perp \mathcal{H}_q.$$

The differential of π at q, $D\pi(q)$, vanishes on $\mathcal{V}_q$ (since π is constant on $\pi^{-1}(m)$) and is an isomorphism between $\mathcal{H}_q$ and T_mM. Let us make the abuse of notation of still denoting $D\pi(q)$ the restriction of $D\pi(q)$ to $\mathcal{H}_q$. Then, if $q, q' \in \pi^{-1}(m)$, the map $p_{q',q} := D\pi(q)^{-1} \circ D\pi(q')$ is an isomorphism between $\mathcal{H}_{q'}$ and $\mathcal{H}_q$. One says that π is a Riemannian submersion if and only if the maps $p_{q',q}$ are in fact isometries between $\mathcal{H}_{q'}$ and $\mathcal{H}_q$ whenever q and q' belong in the same fiber, that is, if one has, for all $v' \in \mathcal{H}_{q'}$,

$$\|p_{q',q}v'\|_q = \|v'\|_{q'}.$$

Another way to phrase this property is

$$\pi(q) = \pi(q'), v \in \mathcal{H}_q, v' \in \mathcal{H}_{q'}, D\pi(q)v = D\pi(q')v' \Rightarrow \|v\|_q = \|v'\|_{q'}.$$

A Riemannian submersion naturally induces a Riemannian metric on M, simply defining, for $m \in M$ and $h \in T_m M$,

$$\|h\|_m = \|D\pi(q)^{-1}h\|_q$$

for any $q \in \pi^{-1}(m)$, the definition being independent of q by assumption. This is the Riemannian projection of the metric on Q via the Riemannian submersion π.

Let us now return to the landmark case, and consider the action of rotations and translations, that is of the special Euclidean group of $\mathbb{R}^d$, which is traditionally denoted $SE(\mathbb{R}^d)$. The action of a transformation $g \in SE(\mathbb{R}^d)$ on a landmark configuration $\mathbf{x} = (x_1, \ldots, x_N) \in Lmk_N$ is

$$g \cdot \mathbf{x} = (g(x_1), \ldots, g(x_N)).$$

We want to use a Riemannian projection to deduce a metric on the quotient space $M = Lmk_N / SE(\mathbb{R}^d)$ from the metric that has been defined on Lmk_N, the surjection π being the projection $\pi : Lmk_N \to M$, which assigns to a landmark configuration $\mathbf{x}$ its equivalence class, or orbit under the action of $SE(\mathbb{R}^d)$, defined by

$$[\mathbf{x}] = \{g \cdot \mathbf{x}, g \in SE(\mathbb{R}^d)\} \in M.$$

To make sure that M is a manifold, one needs to restrict to affinely independent landmark configurations, which form an open subset of Lmk_N and therefore let Q be this space and restrict π to Q. In this context, one can show that a sufficient condition for π to be a Riemannian submersion is that the action of $SE(\mathbb{R}^d)$ is isometric, that is, for all $g \in SE(\mathbb{R}^d)$, the operation $a_g : \mathbf{x} \mapsto g \cdot \mathbf{x}$ is such that, for all $u, v \in T_{\mathbf{x}} Q$,

$$\left\langle Da_g(\mathbf{x})u\,,\, Da_g(\mathbf{x})v \right\rangle_{g \cdot \mathbf{x}} = \langle u\,,\, v \rangle_{\mathbf{x}}.$$

This property can be translated into equivalent properties on the metric. For translations, for example, it says that, for every $\mathbf{x} \in Q$ and $\tau \in \mathbb{R}^d$, one must have

$$S_V(\mathbf{x} + \tau) = S_V(\mathbf{x})$$

which is in turn equivalent to the requirement that, for all $x, y, \tau \in \mathbb{R}^d$, $K_V(x + \tau, y + \tau) = K_V(x, y)$, so that K_V only depends on $x - y$. With rotations, one needs

$$\mathrm{diag}(R)^T S_V(R\mathbf{x})\mathrm{diag}(R) = S_V(\mathbf{x}),$$

which again translates into a similar property for the kernel, namely,

$$R^T K_V(Rx, Ry)R = K_V(x, y).$$

Here, R is an arbitrary d dimensional rotation, and $\mathrm{diag}(R)$ is the dN by dN block-diagonal matrix with R repeated N times.

Kernels that satisfy these properties can be characterized in explicit forms. These kernels include all positive radial kernels, that is, all kernels taking the form

$$K_V(x, y) = \gamma(|x - y|^2)\mathrm{Id}_{\mathbb{R}^d},$$

where $\gamma : [0, +\infty) \to \mathbb{R}$ is the Laplace transform of some positive measure μ, that is,

$$\gamma(t) = \int_0^\infty e^{-ty} d\mu(y).$$

Such functions include Gaussians,

$$\gamma(t^2) = \exp(-t^2/2\sigma^2), \tag{7}$$

Cauchy,

$$\gamma(t^2) = \frac{1}{1 + t^2/\sigma^2}, \tag{8}$$

or Laplacian kernels, defined for any integer $c \geq 0$ by

$$\gamma_c(t^2) = \left(\sum_{l=1}^{c} \rho(c, l)\frac{t^l}{\sigma^l}\right)\exp(-t/\sigma) \tag{9}$$

with $\rho(c, l) = 2^{l-c}(2c - l)\cdots(c + 1 - l)/l!$.

One can also use non-diagonal kernels. One simple construction of such kernels is to start with a scalar kernel, for example, associated to a radial function γ as above, and, for some parameter $\lambda \geq 0$, to implicitly define K_V via the identity, valid for all pairs of smooth compactly supported vector fields V and w,

$$\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} v(x)^T K_V(x, y)w(y)dxdy = \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \gamma\left(|x - y|^2\right)(v(x)^T w(y))dxdy + \frac{\lambda}{2}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \gamma\left(|x - y|^2\right)(\nabla \cdot v(x))(\nabla \cdot w(y))dxdy,$$

where $(\nabla \cdot)$ denotes the divergence operator. The explicit form of the kernel can be deduced after a double application of the divergence theorem yielding

$$K_V(x, y) = (\gamma(r^2) - \lambda\dot{\gamma}(r^2))\mathrm{Id}_{\mathbb{R}^d} - 2\lambda\ddot{\gamma}(r^2)(x - y)(x - y)^T$$

with $r = |x - y|$.

Assume that one of these choices has been made for K_V, so that one can use a Riemannian submersion to define a metric on the quotient space $Q/SE(\mathbb{R}^d)$. One of the appealing properties of this construction is that geodesics in the quotient space are given by (equivalent classes of) geodesics in the original space, provided that they are initialized with horizontal velocities. Another interesting feature is that the horizontality condition is very simply expressed in terms of the momenta, which provides another advantage of the momentum representation in Eq. (6). Take translations, for example. A vertical tangent vector for their action at any point $\mathbf{x} \in M$ is a vector of the form $(\tau, \ldots, \tau)$, where τ is a d-dimensional vector repeated N times. A momentum, or covector, $\boldsymbol{\alpha}$ is horizontal if and only if it vanishes when applied to any such vertical vector, which yields

$$\sum_{k=1}^{N} \alpha_k = 0. \tag{10}$$

A similar analysis for rotations yields the horizontality condition

$$\sum_{k=1}^{N} \left(\alpha_k x_k^T - x_k \alpha_k^T\right) = 0. \tag{11}$$

These two conditions provide the $d(d + 1)/2$ constraints that must be imposed to the momentum representation on M to obtain a momentum representation on $M/SE(\mathbb{R}^d)$.

Hamiltonian Point of View

General Principles

This section presents an alternate formulation in which the accent is made on variational principles rather than on geometric concepts. Although the results obtained using the Hamiltonian approach that is presented here will be partially redundant with the ones that were obtained using the Riemannian point of view, there is a genuine benefit in understanding and being able to connect the two of them. As will be seen below, working with the Hamiltonian formulation brings new, relatively simple concepts, especially when dealing with invariance and symmetries. It is also often the best way to handle numerical implementations.

To lighten the conceptual burden, the presentation will remain within the elementary formulation that uses a state variable q and a momentum p, rather

than the more general symplectic formulation. On a manifold, this implies that the presentation is made with variables restricted to a local chart.

An optimal control problem in Lagrangian form is associated to a real-valued cost function (or Lagrangian) $(q,u) \mapsto L(q,u)$ defined on $Q \times U$, where Q is a manifold and U is the space of controls, and to a function $(q,u) \mapsto f(q,u) \in T_q Q$. The resulting variational problem consists in the minimization of

$$\int_{t_i}^{t_f} L(q,u)dt \tag{12}$$

subject to the constraint $\dot{q}_t = f(q,u)$ and some boundary conditions for $q(t_i)$ and $q(t_f)$. The simplest situation is the classical problem in the calculus of variations for which $f(q,u) = u$, and the problem is to minimize $\int_{t_i}^{t_f} L(q,\dot{q}_t)dt$. Here, $[t_i, t_f]$ is a fixed finite interval. The values $t_i = 0$ and $t_f = 1$ will be assumed in the following.

The general situation in (12) can be formally addressed by introducing Lagrange multipliers, denoted $p(t)$, associated to the constraint $\partial_t q = f(q,u)$ at time t; p is called the costate in the optimal control setting. One then looks for critical paths of

$$J_0(q,p,u) \doteq \int_0^1 \left(L(q,u) + (p|\,\dot{q}_t - f(q,u))\right) dt,$$

where the paths p, u, and q vary now freely as far as $q(0)$ and $q(1)$ remain fixed. The costate is here a linear form on Q, that is, an element of $T_q^* Q$.

Introduce the Hamiltonian

$$H(q,p,u) \doteq (p|\,f(q,u)) - L(q,u)$$

for which

$$J_0 = \int_0^1 \left((p|\,\dot{q}_t) - H(q,p,u)\right) dt.$$

Writing the conditions for criticality, $\delta J_0/\delta u = \delta J_0/\delta q = \delta J_0/\delta p = 0$, directly leads to the Hamiltonian equation:

$$\partial_t q = \partial_p H,\ \partial_t p = -\partial_q H,\ \partial_u H = 0. \tag{13}$$

The above derivation is only formal. A rigorous derivation in various finite-dimensional as well as infinite-dimensional situations is the central object of Pontryagin Maximum Principle (PMP) theorems which state that along a solution (q_*, p_*, u_*), one has

$$H(q_*(t), p_*(t), u_*(t)) = \max_u H(q_*(t), p_*(t), u).$$

Introducing $\tilde{H}(q, p) \doteq \max_u H(q, p, u)$, one gets the usual Hamiltonian equation:

$$\partial_t p = -\partial_q \tilde{H}, \ \partial_t q = \partial_p \tilde{H}. \tag{14}$$

One can notice that, in the classical case $f(q, u) = u$, $\tilde{H}(q, p)$ coincides with the Hamiltonian obtained via the Legendre transformation in which a function $u(p, q)$ is defined via the equation $p = \partial_u L$ and

$$\tilde{H}(p, q) = (p|\, u(q, p)) - L(q, u(q, p)).$$

Application to Geodesics in a Riemannian Manifold

Let Q be a Riemannian manifold with metric at q denoted $\langle \cdot\,,\, \cdot \rangle_q$. The computation of geodesics in Q can be seen as a particular case of the previous framework in at least two (equivalent) ways. The first one is to take

$$L(\mathbf{x}, u) = \|u\|_q^2/2 \text{ and } f(\mathbf{x}, u) = u,$$

which gives a variational problem in standard form. For the other choice, introduce the duality operator $K_q : T_q^* Q \to T_q Q$ defined by

$$(\alpha|\, \xi) = \big\langle K_q \alpha\,,\, \xi \big\rangle_q,$$

$\alpha \in T_q^* Q$, $\xi \in T_q Q$, and let, denoting the control by α,

$$L(q, \alpha) = \big(\alpha|\, K_q \alpha\big)/2 \text{ and } f(q, \alpha) = K_q \alpha.$$

The Hamiltonian equation in this case yields $p = \alpha$ and

$$\begin{cases} \partial_t q = K_q \alpha, \\ \partial_t \alpha = -\dfrac{1}{2} \partial_q \big((\alpha|\, K_q \alpha) \big). \end{cases} \tag{15}$$

This equation obviously reduces to (5) with $q = \mathbf{x}$, $K_q \alpha = S_V(\mathbf{x})\boldsymbol{\alpha}$.

Momentum Map and Conserved Quantities

A central aspect of the Hamiltonian formulation is its ability to turn symmetries into conserved quantities. This directly relates to the Riemannian submersion discussed in section "Invariance."

Consider a Lie group G acting on the state variable $q \in Q$, assuming, for the rest of this section and the next one, an action on the right denoted $(g, q) \to q \cdot g$. Notice that results obtained with a right action immediately translate to left actions, by transforming a left action $(g, q) \mapsto g \cdot q$ into the right action $(g, q) \mapsto g^{-1} \cdot q$. In fact, both right and left actions are encountered in this chapter. The standard notation $T_{\mathrm{id}} G = \mathfrak{G}$ will be used in the following to represent the Lie algebra of G.

By differentiation in the q variable, the action can be extended to the tangent bundle, with notation $(g, v) \to v \cdot g$ for $v \in TQ$. By duality, this induces an action on the costate variable through the equality $(p \cdot g \,|\, v) \doteq \big(p \,|\, v \cdot g^{-1}\big)$. Differentiating again in g at $g = \mathrm{id}_G$ gives the infinitesimal actions on the tangent and cotangent bundles, defined for any $\xi \in \mathfrak{G} \doteq T_{\mathrm{id}_G} G$ by $(\xi, v) \to v \cdot \xi$ and for any $(\xi, p) \to p \cdot \xi$ such that $(p \cdot \xi \,|\, v) + (p \,|\, v \cdot \xi) = 0$, for all $v \in TQ$ and $p \in T^*Q$.

Now, assume that H is G-invariant, that is, $H(q \cdot g, p \cdot g) = H(q, p)$ for any $g \in G$, and define the *momentum map* $(q, p) \to \mathfrak{m}(q, p) \in \mathfrak{G}^*$ by

$$(\mathfrak{m}(q, p) \,|\, \xi) = (p \,|\, q \cdot \xi). \tag{16}$$

Then, one has, along a Hamiltonian trajectory,

$$\partial_t \mathfrak{m}(p, q) = 0, \tag{17}$$

that is, the momentum map is a conserved (vectorial) quantity along the Hamiltonian flow. This result is proved as follows. First notice that if $g(t)$ is a curve in G with $g(0) = \mathrm{id}_G$ and $\dot{g}_t(0) = \xi$, then, if H is G-invariant,

$$0 = \partial_t H(q \cdot g, p \cdot g) = \big(\partial_q H \,|\, q \cdot \xi\big) + \big(p \cdot \xi \,|\, \partial_p H\big).$$

On the other hand, from the definitions of the actions, one has

$$(\partial_t \mathfrak{m}(q, p) \,|\, \xi) = \partial_t (p \,|\, q \cdot \xi) = (\partial_t p \,|\, q \cdot \xi) - (p \cdot \xi \,|\, \partial_t q),$$

so that if (q, p) is a Hamiltonian trajectory,

$$(\partial_t \mathfrak{m}(q, p) \,|\, \xi) = -\big(\partial_q H \,|\, q \cdot \xi\big) - \big(p \cdot \xi \,|\, \partial_p H\big) = 0$$

which gives (17).

Notice that the momentum map has an interesting equivariance property:

$$\begin{aligned}(\mathfrak{m}(q \cdot g, p \cdot g) \,|\, \xi) &= (p \cdot g \,|\, (q \cdot g) \cdot \xi) \\ &= (p \cdot g \,|\, q \cdot (g\xi)) \\ &= \big(p \,|\, (q \cdot (g\xi)) \cdot g^{-1}\big) \\ &= \big(p \,|\, q \cdot ((g\xi) g^{-1})\big)\end{aligned}$$

where $g\xi$ denotes the derivative of $h \mapsto gh$ in h at $h = \mathrm{id}_G$ along the direction ξ and $(g\xi)g^{-1}$ the derivative of $h \mapsto hg^{-1}$ in h at $h = g$ along the direction $g\xi$. The map $\xi \mapsto (g\xi)g^{-1}$ defined on $\mathfrak{G}$ is called the *adjoint representation* and usually denoted $v \mapsto \mathrm{Ad}_g \xi$. One therefore gets

$$(\mathfrak{m}(q \cdot g, p \cdot g) \,|\, \xi) = \big(p \,|\, q \cdot \mathrm{Ad}_g(\xi)\big) = \big(\mathfrak{m}(q, p) \,|\, \mathrm{Ad}_g(\xi)\big) = \Big(\mathrm{Ad}_g^*(\mathfrak{m}(q, p)) \,|\, \xi\Big),$$

where Ad^*_g is the conjugate of Ad_g. Hence

$$\mathfrak{m}(q \cdot g, p \cdot g) = \mathrm{Ad}^*_g(\mathfrak{m}(q, p)), \tag{18}$$

that is, $\mathfrak{m}$ is Ad^*-equivariant.

Euler–Poincaré Equation

Consider the particular case in which $Q = G$ and G acts on itself. In this case,

$$(\mathfrak{m}(\mathrm{id}_G, p) | v) = (p | v),$$

so that $\mathfrak{m}(\mathrm{id}_G, p) = p$ and one gets from Eq. (18)

$$pg^{-1} = \mathfrak{m}(\mathrm{id}_G, pg^{-1}) = Ad^*_{g^{-1}}(\mathfrak{m}(g, p)).$$

Hence, along a trajectory starting from $g(0) = \mathrm{id}_G$ of a G-invariant Hamiltonian H, one has (denoting $\rho = pg^{-1} \in \mathfrak{G}^*$ and using the fact that the momentum map is conserved over time)

$$\begin{aligned} \rho(t) \doteq p(t)g(t)^{-1} &= \mathrm{Ad}^*_{g^{-1}(t)}(\mathfrak{m}(g(t), p(t))) \\ &= \mathrm{Ad}^*_{g^{-1}(t)}(\mathfrak{m}(\mathrm{id}_G, p(0))) = \mathrm{Ad}^*_{g^{-1}(t)}(\rho(0)). \end{aligned} \tag{19}$$

This is the integrated version of the so-called *Euler–Poincaré* equation on $\mathfrak{G}^*$ [6,50],

$$\partial_t \rho + \mathrm{ad}^*_{v(\rho)}(\rho) = 0, \tag{20}$$

where $v(\rho) = \dot{g}g^{-1} = \partial_p H(\mathrm{id}_G, pg^{-1}) = \partial_p H(\mathrm{id}_G, \rho)$ and ad is the differential at location $g = \mathrm{id}_G$ of Ad_g.

A special case of this, which will be important later, is when the Hamiltonian corresponds to a right-invariant Riemannian metric on G. There is a large literature on invariant metrics on Lie groups, which can be shown to be related to important finite and infinite-dimensional mechanical models, including the Euler equation for perfect fluids. The interested reader can refer to [5,6,33,34,49,50].

Such a metric is characterized by an inner product $\langle \cdot\,,\, \cdot \rangle_V$ on $\mathfrak{G}$ and defined by

$$\langle v\,,\, w \rangle_g = \left\langle vg^{-1}\,,\, wg^{-1} \right\rangle_{\mathfrak{G}}. \tag{21}$$

If one lets $K_{\mathfrak{G}}$ be the duality operator on $\mathfrak{G}$ so that

$$(\rho | v) = \langle K_{\mathfrak{G}} \rho\,,\, v \rangle_{\mathfrak{G}},$$

the issue of finding minimizing geodesics can be rephrased as an optimal control problem like in the case of landmarks, with Lagrangian $L(g, \mu) = \big(\mu | K_g \mu\big)/2$, $f(g, \mu) = K_g \mu$, and

$$K_g\mu = (K_{\mathfrak{G}}(\mu g^{-1}))g. \tag{22}$$

The Hamiltonian equations are then directly given by (15), namely,

$$\begin{cases} \partial_t g = K_g\mu, \\ \partial_t \mu = -\dfrac{1}{2}\partial_g\big((\mu|\, K_g\mu)\big). \end{cases} \tag{23}$$

This equation is equivalent to the one obtained from the conservation of the momentum map, which is (with $\rho = \mu g^{-1}$)

$$\begin{cases} \partial_t g = vg, \\ v = K_{\mathfrak{G}}\rho, \\ \partial_t \rho = -\mathrm{ad}^*_v\rho. \end{cases} \tag{24}$$

A Note on Left Actions

Invariance with respect to left actions is handled in a symmetrical way to right actions. If G is acting on the left on G, define the momentum map by

$$(\mathfrak{m}(p,q)|\, v) = (p|\, v\cdot q)$$

which is conserved along Hamiltonian trajectories. Working out the equivariance property gives

$$\mathfrak{m}(g\cdot p, g\cdot q) = \mathrm{Ad}^*_{g^{-1}}\mathfrak{m}(p,q)\,.$$

When G acts on itself on the left, the Euler–Poincaré equation reads

$$\rho(t) = \mathrm{Ad}^*_g(\rho(0))$$

or

$$\partial_t\rho - \mathrm{ad}^*_{v(\rho)}\rho = 0$$

with $\rho = g^{-1}p$ and $v(\rho) = g^{-1}\dot{g}_t$.

Application to the Group of Diffeomorphisms

Let $G \subset \mathrm{Diff}(\mathbb{R}^d)$ be a group of smooth diffeomorphisms of $\mathbb{R}^d$ (which, say, smoothly converge to the identity at infinity). Elements of the tangent space to G, which are derivatives of curves $t \mapsto \varphi(t,\cdot)$ where $\varphi(t,\cdot) \in G$ for all t, can be identified to vector fields $x \mapsto v(x) \in \mathbb{R}^d$, $x \in \mathbb{R}^d$.

To define a right-invariant metric on G, introduce a Hilbert space V of vector fields on $\mathbb{R}^d$ with inner product $\langle\cdot\,,\,\cdot\rangle_V$. Like in section "A Riemannian Manifold

of Deformable Landmarks," let L_V and $K_V = L_V^{-1}$ denote the duality operators on V, with $\langle v\,,\, w\rangle_V = (L_V v|\, w)$ and $\langle \mu\,,\, \nu\rangle_{V^*} = (\mu|\, K_V \nu)$; K_V is furthermore identified with a matrix-valued kernel $K_V(x, y)$ acting on vector fields.

The application of the formulae derived for Hamiltonian systems and of the Euler–Poincaré equation will remain in the following of this section at a highly formal level, just computing the expression assumed in the case of diffeomorphisms by the general quantities introduced in the previous section. There will be no attempt at proving that these formulae are indeed valid in this infinite-dimensional context, which is out of the scope of this chapter. As an example of the difficulties that can be encountered, let us mention the dilemma that is involved in the mere choice of the group G. On the first hand, G can be chosen as a group of infinitely differentiable diffeomorphisms that coincide with the identity outside a compact set. This would provide a rather nicely behaved manifold with a Lie group structure in the sense of [44, 45]. The problem with such a choice is that the structure would be much stronger than what Riemannian metrics of interest would induce and that geodesics would typically spring out of the group. One can, on the other hand, try to place the emphasis on the Riemannian and variational aspects so that the computation of geodesics in G, for example, remains well posed. This leads to a solution, introduced in Trouvé (Infinite dimensional group action and pattern recognition. Technical report. DMI, Ecole Normale Supérieure, unpublished, 1995) (see also [70]), in which G is completed in a way which depends on the metric $\langle \cdot\,,\, \cdot\rangle_V$, so that the resulting group (denote it G_V) is complete for the geodesic distance. This extension, however, comes with the cost of losing the nice features of infinitely differentiable transformations, resulting in G_V not being a Lie group, for example.

This being acknowledged, first consider the transcription of (23) to the case of diffeomorphisms. This equation will involve a time-evolving diffeomorphism $\varphi(t, \cdot)$, and a time-evolving covector, denoted $\mu(t)$, which is a linear form over vector fields (it takes a vector field $x \mapsto v(x)$ and returns a number that has so far been denoted $(\mu(t)|\, v)$). It will be useful to apply $\mu(t)$ to vector-valued functions of several variables, say $v(x, y)$ defined for $x, y \in \mathbb{R}^2$, by letting one of the variables fixed and considering V as a function of the other. This will be denoted by adding a subscript representing the effective variable, so that

$$(\mu(t)|\, v(x, y))_x$$

is the number, dependent of y, obtained by applying $\mu(t)$ to the vector field $x \mapsto v(x, y)$.

One first needs to identify the operator K_φ in Eq. (22), defined by

$$K_\varphi \mu = (K_V(\mu\varphi^{-1}))\varphi = (K_V(\mu\varphi^{-1})) \circ \varphi$$

since right translation here coincides with composition. Now, for any vector $a \in \mathbb{R}^d$ and $y \in \mathbb{R}^d$, one has

$$\begin{aligned} a^T(K_\varphi \mu)(y) &= a^T(K_V(\mu\varphi^{-1}))(\varphi(y)) \\ &= \big(a \otimes \delta_{\varphi(y)} |\, K_V(\mu\varphi^{-1})\big) \\ &= \big(\mu\varphi^{-1} |\, K_V(a \otimes \delta_{\varphi(y)})\big) \\ &= \big(\mu |\, K_V(a \otimes \delta_{\varphi(y)}) \circ \varphi\big) \\ &= (\mu |\, K_V(\varphi(x), \varphi(y))a)_x. \end{aligned}$$

So, letting $e_1, \dots, e_d$ denote the canonical basis of $\mathbb{R}^d$, one has

$$(K_\varphi \mu)(y) = \sum_{i=1}^{d} e_i^T (K_\varphi \mu)(y)\, e_i = \sum_{i=1}^{d} \big(\mu |\, K_V^i(\varphi(x), \varphi(y))\big)_x e_i,$$

where K_V^i is the ith column of K_V. Therefore

$$\big(\mu |\, K_\varphi \mu\big) = \sum_{i=1}^{d} \big(\mu |\, \big(\mu |\, K_V^i(\varphi(x), \varphi(y))\big)_x e_i\big)_y$$

and (using the symmetry of K_V)

$$\big(\partial_\varphi\big(\mu |\, K_\varphi \mu\big) |\, w\big) = 2 \sum_{i=1}^{d} \big(\mu |\, \big(\mu |\, D_2 K_V^i(\varphi(x), \varphi(y))w(y)\big)_x e_i\big)_y,$$

where $D_2 K_V^i$ is the derivative of K_V with respect to its second variable. These computations directly give the transcription of (23) for diffeomorphisms, namely,

$$\begin{cases} \partial_t \varphi(t, y) = \displaystyle\sum_{i=1}^{d} \big(\mu(t) |\, K_V^i(\varphi(t, x), \varphi(t, y))\big)_x e_i \\ \forall w : (\partial_t \mu(t) |\, w) = -\displaystyle\sum_{i=1}^{d} \big(\mu(t) |\, \big(\mu(t) |\, D_2 K_V^i(\varphi(t, x), \varphi(t, y))w(y)\big)_x e_i\big)_y. \end{cases} \tag{25}$$

To transcribe Eq. (24) to diffeomorphisms, one only needs to work out the expressions of Ad_φ and ad_v in this context. Recall that $\mathrm{Ad}_\varphi w$ was defined by $(\varphi w)\varphi^{-1}$; φw being the differential of the left translation (i.e., $\partial_t(\varphi \circ \psi(t))(0)$ with $\psi(0) = \mathrm{id}$ and $\partial_t \psi(0) = w$), one finds $\varphi w = D\varphi\, w$, and since right translation is just composition,

$$\mathrm{Ad}_\varphi w = (D\varphi\, w) \circ \varphi^{-1}.$$

Now, since $\mathrm{ad}_v w$ is the differential of $\mathrm{Ad}_\varphi w$ in φ (at $\varphi = \mathrm{id}$), a quick computation shows that

$$\mathrm{ad}_v w = Dv\, w - Dw\, v.$$

So, Eq. (24) provides

$$\begin{cases} \partial_t \varphi(t,y) = v(t,\varphi(t,y)) \\ v(t,x) = K_V \rho(t)(x) \\ \forall w \in V, (\partial_t \rho \,|\, w) = -(\rho(t) \,|\, Dv\, w - Dw\, v) \end{cases} \tag{26}$$

with the last equation being equivalent to

$$(\rho(t) \,|\, w) = \big(\rho(0) \,|\, D\varphi^{-1}(w \circ \varphi)\big).$$

Note that μ and ρ in (25) and (26) are related via $\mu = \rho\varphi$ or

$$(\mu \,|\, w) = \big(\rho \,|\, w \circ \varphi^{-1}\big).$$

Solving Eq. (25) (or (26)) between times 0 and 1 provides the momentum representation in G, denoted

$$\varphi(1,\cdot) = \exp^{\flat}_{\varphi(0,\cdot)}(\mu(0)).$$

Equivalently, the initial velocity being $K_V \mu(0)$, this is, in the exponential chart,

$$\varphi(1,\cdot) = \exp_{\varphi(0,\cdot)}(K_V \mu(0)).$$

Reduction via a Submersion

This section, which can be put in parallel with the discussion on Riemannian submersions in section "Invariance," discusses how submersions from a manifold Q onto another manifold M allow for the transfer of a Hamiltonian system on Q to a Hamiltonian system on M, given some invariance properties satisfied by the Hamiltonian.

Let π be a submersion from a manifold Q onto a manifold M so that for any $q \in Q$, $D\pi_q : T_qQ \to T_{\pi(q)}M$ is a surjective mapping. For any $q \in Q$, $\mathcal{V}_q \doteq D\pi(q)^{-1}(0)$ is the previously mentioned vertical space so that $\mathcal{V} \doteq \cup_{q \in Q} \mathcal{V}_q$ will be called the vertical bundle. In the previous Riemannian setting, a metric on TQ induces the definition of a horizontal space $\mathcal{H}_q$ at any location $q \in Q$ such that $T_qQ = \mathcal{V}_q + \mathcal{H}_q$. In the Hamiltonian approach, the horizontal space is defined in the cotangent space T^*_qQ *without any reference to a particular metric* as the set of conormal covectors to the vertical space, that is,

$$\mathcal{H}^*_q \doteq \left\{ p \in T^*_qQ \mid (p \,|\, v) = 0 \ \forall v \in \mathcal{V}_q \right\}. \tag{27}$$

An elementary argument in linear algebra (which is left to the reader) shows that if one introduces the one-to-one adjoint mapping $D\pi^*(q) : T^*_{\pi(q)}M \to T^*_qQ$, one has

$\mathcal{H}_q^* = D\pi(\pi(q))^*\big(T^*_{\pi(q)}M\big)$. In other terms, a covector p is horizontal at q if and only if there exists a covector $\alpha \in T^*_{\pi(q)}M$ such that $D\pi(\pi(q))^*\alpha = p$. Therefore, $\mathcal{H}^* \doteq \cup_{q\in Q}\mathcal{H}_q^*$ can be seen as a sub-bundle of the cotangent bundle T^*Q for which there exists a surjective mapping

$$\tilde{\pi} : \mathcal{H}^* \to T^*M$$

defined by $\tilde{\pi}(q, p) = (\pi(q), (D\pi^*(q))^{-1}(p))$.

The main idea of this Hamiltonian (one should say symplectic or better Poisson) point of view is that $\mathcal{H}^*$ is the natural image in T^*Q of the dynamic space (phase space) T^*M on M. Now, assume that a Hamiltonian H_Q is given on Q. One says that H_Q is π-reducible if there exists a Hamiltonian H_M on T^*M such that

$$H_{Q\,|\mathcal{H}^*} = H_M \circ \tilde{\pi} \tag{28}$$

or equivalently

$$H_M(m, \alpha) = H_Q(q, D\pi(q)^*\alpha) \tag{29}$$

for $q \in \pi^{-1}(m)$.

Hamiltonian trajectories in both spaces are related as follows. Assume that (q, p) is H_Q-Hamiltonian (i.e., $\partial_t q = \partial_p H_Q$ and $\partial_t p = -\partial_q H_Q$) with $(q(0), p(0)) \in \mathcal{H}^*$, and consider in a similar way a H_M-Hamiltonian trajectory (m, α) such that $(m(0), \alpha(0)) = \tilde{\pi}(q(0), p(0))$; then for any $t \geq 0$, one has $(q(t), p(t)) \in \mathcal{H}^*$ and $(m(t), \alpha(t)) = \tilde{\pi}(q(t), p(t))$. Equivalently, one has the commutative diagram

$$\begin{array}{ccc} \mathcal{H}^* & \xrightarrow{\Phi_Q(\cdot,\cdot,t)} & \mathcal{H}^* \\ \big\downarrow{\scriptstyle \tilde{\pi}} & & \big\downarrow{\scriptstyle \tilde{\pi}} \\ T^*M & \xrightarrow{\Phi_M(\cdot,\cdot,t)} & T^*M, \end{array} \tag{30}$$

where Φ_H and Φ_Q are the associated Hamiltonian flows (in particular $\mathcal{H}^*$ is Φ_Q invariant). To prove this fact, first, notice that from the definition of H_M, which can be rewritten as

$$H_M(\pi(q), \alpha) = H_Q(q, D\pi(q)^*\alpha),$$

one gets

$$(\rho|\,\partial_\alpha H_M) = \big(D\pi(q)^*\rho|\,\partial_p H_Q\big) \tag{31}$$

$$\text{and } (\partial_m H_M|\,D\pi(q)\xi) = \big(\partial_q H_Q|\,\xi\big) + \big(\alpha|\,D^2\pi(q)(\xi, \partial_p H_Q)\big) \tag{32}$$

(as usual, computations are assumed to be done within a chart, and the second derivative of π is defined according to this chart).

Define $x(t) \doteq (m(t), \alpha(t))$, $y(t) \doteq (q(t), p(t)))$, $z = (x, y)$ and the transformation

$$\psi(z) = (\psi_M(z), \psi^Q(z)) = (m - \pi(q), p - D\pi^*(q)\alpha).$$

One needs to prove that $\psi(z(t)) \equiv 0$.

If $Z(z) \doteq (\partial_\alpha H_M, -\partial_m H_M, \partial_p H_Q, -\partial_q H_Q)$ is the vector field governing the joint Hamiltonian flows (by construction $\partial_t z = Z(z)$), one has

$$D\psi(z)Z = 0 \text{ if } \psi(z) = 0. \tag{33}$$

Notice that this fact implies that Z is everywhere tangent to the set $\psi = 0$, which is locally a submanifold because $D\psi(z)$ has full rank, as can easily be seen. This implies that $\psi = 0$ is invariant by the flow associated to Z.

To prove (33), assume $\psi(z) = 0$ and notice that the statement is equivalent to $(\zeta^M | D\psi_M(z)Z(z)) + (D\psi^Q(z)Z(z)|\zeta_Q) = 0$ for any $\zeta = (\zeta^M, \zeta_Q)$. One has

$$(\zeta^M | D\psi_M(z)Z(z)) = (\zeta^M | \partial_\alpha H_M - D\pi(q)\partial_p H_Q)$$

and

$$\begin{aligned}\big(D\psi^Q(z)Z(z)|\,\zeta_Q\big) &= \big(-\partial_q H_Q + D\pi^*(q)\partial_m H_M\,|\,\zeta_Q\big) - \big(\alpha|\,D^2\pi(q)(\zeta_Q, \partial_p H)\big) \\ &= \big(\partial_m H_M|\,D\pi(q)\zeta_Q\big) - \big(\partial_q H_Q|\,\zeta_Q\big) - \big(\alpha|\,D^2\pi(q)(\partial_p H, \zeta_Q)\big),\end{aligned}$$

and the result is a direct consequence of (31) and (32).

Let us review how this concept of reduction via a submersion property generalizes the Riemannian submersion idea. When Q and M are Riemannian with M equipped with the projected metric, one has by construction

$$\langle \xi\,,\,\eta\rangle_q = \langle D\pi(q)\xi, \quad D\pi(q)\eta\rangle_{\pi(q')} \tag{34}$$

for any $q \in M$ and $\xi \in \mathcal{H}_q$ horizontal at q. Let $K_q : T_q^*Q \to T_qQ$ be the duality operator for the metric at q (such that $(p|\,\xi) = \big\langle K_q p\,,\,\xi\big\rangle_q$) and K_m be the same operator for the metric on M. From (27), one gets

$$\mathcal{H}_q = K_q \mathcal{H}_q^*\,.$$

If one expresses (34) for $\xi = K_q D\pi(q)^*\alpha$ and $\eta = K_q D\pi(q)^*\beta$ and identifies the terms, one gets an equivalent version of (34) in terms of the duality operators, namely,

$$K_m = D\pi(q)K_q D\pi(q)^*, \tag{35}$$

the invariance assumption being that the right-hand term does not depend on $q \in \pi^{-1}(m)$. Now, the Hamiltonians associated to the metrics, respectively, are $H_Q(q,p) = \big(p|\, K_q\, p\big)/2$ and $H_M(m,\alpha) = (\alpha|\, K_m\alpha)/2$, and it is now straightforward to see that the condition $H_M(m,\alpha) = H_Q(q, D\pi(q)^*\alpha)$ if $\pi(q) = m$, that is, condition (28), is also equivalent to (35).

Reduction: Quotient Spaces

A fundamental special case of the previous situation is when π is the projection onto a quotient space $M = Q/G_s$ where G_s is a group of symmetries, acting on Q. A left action is assumed in the following, a right action being handled in a symmetrical way. Introduce the canonical projection $\pi : Q \to M$ which associates the orbit $G_s \cdot q$ to an element q of Q. Let us first work out conditions that ensure that a Hamiltonian H_Q is π-reducible. One needs

$$H_Q(q,p) = H_Q(q',p')$$

whenever $\pi(q) = \pi(q')$, and there exists $\alpha \in T^*_{\pi(q)}M$ such that $p = D\pi(q)^*\alpha$ and $p' = D\pi(q')^*\alpha$. Notice that $\pi(q) = \pi(q')$ implies that there exists a $g \in G$ such that $q' = g \cdot q$. From the relation $D\pi(q')(g \cdot \xi) = D\pi(q)\xi$ which derives from $\pi(g \cdot q) = \pi(q)$, one gets

$$(p|\, \xi) = (\alpha|\, D\pi(q)\xi) = \big(\alpha|\, D\pi(q')(g \cdot \xi)\big) = \big(p'|\, g \cdot \xi\big)$$

which implies that $p' = g \cdot p$ (this condition obviously implying that they correspond to the same α if they both are horizontal). So H_Q is π-reducible if and only if H_Q is G-invariant, namely,

$$H_Q(g \cdot q, g \cdot p) = H_Q(q,p) \tag{36}$$

For the construction made in the previous section to be useful in practice, one needs to provide a simple description of the cotangent bundle to M, T^*M. This will be done using the momentum map $\mathfrak{m}_s$ for the action of G_s, and in particular the set

$$\mathfrak{m}_s^{-1}(0) = \{(q,p) \in T^*Q : \forall \xi \in \mathfrak{G}_s, (p|\, \xi \cdot q) = 0\} = \mathcal{H}^*.$$

Given this notation, one has the identification

$$\mathfrak{m}_s^{-1}(0)/G_s \cong T^*M. \tag{37}$$

First, notice that the right-hand term is meaningful, since, by the equivariance of the momentum map, $\mathfrak{m}_s^{-1}(0)$ is invariant by G_s. To prove (37), recall the transformation $\tilde{\pi} : \mathcal{H}^* = \mathfrak{m}_s^{-1}(0) \to T^*M$ by $\tilde{\pi}(q,p) = (m,\alpha)$ with $m = \pi(\alpha)$ and $p = D\pi(q)^*\alpha$. The last identity means that

$$(\alpha|\, D\pi(q)v) = (p|\, v),$$

and the condition $\mathfrak{m}_s(q, p) = 0$ implies that this definition is not ambiguous, since $D\pi(q)v = 0$ implies that $v = \xi \cdot q$ for some ξ, and therefore that $(p|\, v) = (\mathfrak{m}_s(q, p)|\, \xi) = 0$. (The definition does define $(\alpha|\, \rho)$ for all ρ because $D\pi(q)$ has full rank, since π is a submersion.)

The next remark is that $\tilde{\pi}$ induces a map $[\tilde{\pi}]$ on the quotient space $\mathfrak{m}_s^{-1}(0)/G_s$, defined by

$$[\tilde{\pi}\,](G_s \cdot (q, p)) = \tilde{\pi}(q, p).$$

Again, one must make sure that the definition makes sense by proving that $\tilde{\pi}(g \cdot q, g \cdot p) = \tilde{\pi}(p, q)$, but this is an immediate consequence of the definition of the extended action of G_s on T^*Q. Finally, $[\tilde{\pi}]$ is one to one, since as shown above if $\pi(q) = \pi(q') = m$ and $p = D\pi^*(q)\alpha$ and $p' = D\pi^*(q')\alpha$, then there exists $g \in G_s$ such that $(q', p') = (g \cdot q, g \cdot p)$. This proves the identification (37).

As an example, consider the reduction of the Hamiltonian

$$H(\mathbf{x}, \boldsymbol{\alpha}) = \frac{1}{2}\boldsymbol{\alpha}^T S_V(\mathbf{x})\boldsymbol{\alpha}$$

in the landmark case ($Q = Lmk_N$) and the invariance by the group $G_s = SE(\mathbb{R}^d)$. With $(\boldsymbol{\alpha}|\, \xi) = \sum_{k=1}^N \alpha_k^T \xi_k$, the momentum map for this action is

$$(\mathbf{m}_s(\mathbf{x}, \alpha)|\, (A, \tau)) = \sum_{k=1}^N \alpha_k^T (Ax_k + \tau)$$

defined for all skew-symmetric matrix A and vector $\tau \in \mathbb{R}^d$, and the conditions for $\mathfrak{m}_s(\mathbf{x}, \boldsymbol{\alpha}) = 0$ are exactly those given in (10) and (11).

Note that condition (35) on the duality operator directly corresponds to the invariance conditions associated to the kernel K_V in section "Invariance."

Reduction: Transitive Group Action

Consider the situation of a *left* group action $G \times M \to M$ of a group G on a manifold M. The important example in this chapter is when G is a group of diffeomorphisms and M is a set of "shapes" (for instance, $M = Lmk_N$). Assume that the action is transitive, that is, $G \cdot m_0 = M$ so that $\pi : G \to M$ defined by $\pi(g) = g \cdot m_0$ is a smooth surjection, that will be assumed to be a submersion. The situation here is on how to project a Hamiltonian system on G onto a reduced one on M.

Let

$$G_0 = \{g \in G \mid g \cdot m_0 = m_0\} = \pi^{-1}(m_0)$$

be the isotropy group of m_0. Then condition (29) for a Hamiltonian H_G on G is equivalent to the invariance of H_G to the *right* action of G_0 on G, namely, $H_G(gh, \rho h) = H_G(g, \rho)$ for $h \in G_0$.

Although it is often more convenient to apply the reduction directly to π as defined above, since the structure of T^*M is generally easily defined in this context, it is interesting to notice that this reduction also comes as an application of the previous construction on quotient spaces via the well-known identification [32] $M \cong G/G_0$. This identity extends to cotangent spaces as above, with

$$\mathfrak{m}_G^{-1}(0)/G_0 \cong T^*M, \tag{38}$$

where $\mathfrak{m}_G$ is the momentum map associated with G_0.

One can interpret the construction of the Riemannian metric for landmarks within this framework. Take $M = Lmk_N$, G a group of diffeomorphisms, and $m_0 = \mathbf{x}_0$. If $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_N) \in T_{\mathbf{x}}M^*$, one can identify $p = D\pi^*(\varphi)\alpha$ as

$$p = \sum_{i=1}^{N} \alpha_i \otimes \delta_{x_{0,i}},$$

since for any $v \in T_\varphi G$,

$$(p|\, v) = \sum_{i=1}^{N} (\alpha_i | v(x_{0,i}))$$

and $D\pi(\varphi)v = (v(x_{0,1}), \ldots, v(x_{0,N}))$. Assume that a Riemannian metric is defined on G such that $\langle v\,,\, w\rangle_\varphi$ is associated with a duality operator K_φ that can be identified with a reproducing kernel also denoted K_φ (without assuming right invariance yet). With this assumption, one has

$$H_G(\mathbf{x}, \boldsymbol{\alpha}) = H(\varphi, p) = \frac{1}{2}(p|K_\varphi p) = \frac{1}{2}\sum_{i=1}^{N} \alpha_i^T K_\varphi(x_{0,i}, x_{0,j})\alpha_j$$

The invariance assumption is now clear: one needs that $K_\varphi(x_{0,i}, x_{0,j})$ only depends on $\mathbf{x} = \varphi \cdot \mathbf{x}_0$. This is in particular implied by the full right-invariance assumption discussed in section "Application to the Group of Diffeomorphisms" for which $K_\varphi(x_{0,i}, x_{0,j}) = K_V(x_i, x_j)$, yielding in this case

$$H_M(\mathbf{x}, \boldsymbol{\alpha}) = \frac{1}{2}\sum_{i=1}^{N} \alpha_i^T K_V(x_i, x_j)\alpha_j$$

in the G-invariant case. As an alternative, one could, for example, also use the less restrictive assumption $K_\varphi(x_{0,i}, x_{0,j}) = K_{\mathbf{x}}(x_i, x_j)$ where $K_{\mathbf{x}}$ is still a kernel, like

in (7), (8), or (9), in which the scale parameter σ is chosen dependent of $\mathbf{x}$ (e.g., increasing as a function of $|\mathbf{x} - \mathbf{x}_0|^2$).

The situation of a fully G-invariant Hamiltonian H_G can be studied in the general setting. Indeed, since G acts on M, one can consider the associated momentum map $\mathfrak{m}_M$ on T^*M defined by

$$(\mathfrak{m}_M(m,\alpha)|\,\xi) = (\alpha|\,\xi \cdot m).$$

If $p = D\pi^*(g)(\alpha)$, then $pg^{-1} = \mathfrak{m}_M(m,\alpha)$. Indeed,

$$(pg^{-1}|\xi) = (p|\xi g) = (\alpha|D\pi(g)\xi g) = (\alpha|\xi \cdot m).$$

Hence,

$$H_M(m,\alpha) = H_G(g,p) = H_G(\mathrm{id}_G, pg^{-1}) = H_G(\mathrm{id}_G, \mathfrak{m}_M(m,\alpha)).$$

In the case of an invariant Riemannian metric $H(\mathrm{id}_G, p) = \frac{1}{2}(p|K_V p) = \frac{1}{2}\|p\|^2_{\mathfrak{G}^*}$ where $\|\ \|_{\mathfrak{G}^*}$ denotes the dual norm, this gives

$$H_M(m,\alpha) = \frac{1}{2}\|\mathfrak{m}_M(m,\alpha)\|^2_{\mathfrak{G}^*}\,. \tag{39}$$

Spaces of Plane Curves

Introduction and Notation

We now consider two-dimensional shapes represented by their contours and address the case of spaces of plane curves. Compared to the space of landmarks, two new issues significantly complicate the theory. The first one is that curves are infinite-dimensional objects, which will place us in the framework of infinite-dimensional Riemannian manifolds. The second one is that curves are rarely labeled, which will require the analysis to be invariant by a change of parameterization.

Let us first start with a few definitions regarding plane curves. Parameterized plane curves can be seen as functions $\mathbf{x} : S^1 \to \mathbb{R}^2$, where S^1 is the unit circle in $\mathbb{R}^2$. For simplicity, they will be assumed to be smooth (infinitely differentiable), unless specified otherwise. Smooth curves over the unit circle can equivalently be seen as infinitely differentiable 2π-periodic functions with periodic derivatives defined on the real line. It will be convenient to use both representations in the following.

One says that $\mathbf{x} : S^1 \to \mathbb{R}^2$ is an immersion (or an immersed curve) if its first differential never vanishes (one often also says that $\mathbf{x}$ is a regular curve). We let $\mathcal{I}$ denote the space of immersed curves. Immersed curves, which are easily characterized by their non-vanishing first derivative, are a convenient but a relatively imperfect representation of two-dimensional shapes, since they may include curves that self-intersect. A more restrictive class is the space of embedded curves, that contains immersed curves that coincide, in the neighborhood of any point, and

after a suitable rotation, with the graph of a smooth function. But because being embedded is a global statement about the curve, and therefore harder to handle than being immersed which is just local, this discussion will primarily focus on the space $\mathcal{I}$.

We let $\boldsymbol{\tau}(u) = \dot{\mathbf{x}}(u)/|\dot{\mathbf{x}}(u)|$ be the unit tangent at u (or $\mathbf{x}(u)$) to $\mathbf{x}$; $\boldsymbol{\nu}(u)$ be the unit normal, obtained by rotating $\boldsymbol{\tau}(u)$ of $\pi/2$; and $\boldsymbol{\kappa}(u)$ be the curvature, given by

$$\boldsymbol{\kappa} = (\partial_s \boldsymbol{\tau})^T \boldsymbol{\nu} = \dot{\boldsymbol{\tau}}_u^T \boldsymbol{\nu}/|\dot{\mathbf{x}}_u|$$

where, following [55], we let ∂_s denote the operator $\partial_u/|\dot{\mathbf{x}}_u|$.

A change of parameter (or reparameterization) for a curve is a smooth diffeomorphism $u \mapsto \psi(u)$ of S^1 or, alternatively, a smooth increasing diffeomorphism of the real line such that for all $u \in \mathbb{R}$,

$$\psi(u + 2\pi) = \psi(u) + 2\pi.$$

Changes of parameter act on parameterized curves on the right via

$$(\psi, \mathbf{x}) \mapsto \mathbf{x} \circ \psi.$$

A normalized arc-length parameterization of $\mathbf{x}$ is a change of parameter taking the form

$$s(u) = s_0 + \frac{2\pi}{L} \int_0^u |\dot{\mathbf{x}}_u(\tilde{u})| d\tilde{u} \tag{40}$$

and L is the length of $\mathbf{x}$ with

$$\text{length}(\mathbf{x}) = \int_0^{2\pi} |\dot{\mathbf{x}}_u(\tilde{u})| d\tilde{u}.$$

The scalar number s_0 intervening in the arc-length parameterization can be assumed to be between 0 and 2π without loss of generality and will be referred to as the offset of the parameterization.

The quotient space of immersed curves by reparameterization is the space of geometric curves, denoted $\mathcal{B}$. This space can in turn be quotiented out by the actions of rotations, translations, and scaling, which act on the left and commute with changes of parameter, in the sense that the result of applying a similitude and a change of parameters does not depend on the order with which these operations are performed.

The goal in this section is to discuss shape spaces of curves obtained by putting a Riemannian structure on $\mathcal{B}$, possibly quotiented by Euclidean transformations and/or scaling. But before this discussion, it will be interesting to list a few of the basic distances that can be defined on this set without using a Riemannian construction.

Some Simple Distances

We here consider some simple parameterization-free distances between curves based on the images of the curves (the set $\mathbf{x}(\mathbb{R})$).

A very simple example is to use standard norms (like L^p or Sobolev norms) computed on the difference between two curves parameterized with their normalized arc length. Take, for example, the L^2 norm, and define for two curves $\mathbf{x}$ and $\tilde{\mathbf{x}}$ parameterized with normalized arc length

$$d_{L^2}(\mathbf{x}, \tilde{\mathbf{x}}) = \inf_{s_0} \left(\int_0^{2\pi} |\mathbf{x}(s + s_0) - \tilde{\mathbf{x}}(s)|^2 ds \right) \tag{41}$$

the infimum being taken over all possible offsets as defined in Eq. (40).

One must apply some care when defining distances like (41) which involves some optimization over some parameters that affect the curves. The following statement (the proof of which is left to the reader) is a key for this to be a valid way of building distances on quotient spaces.

Lemma 1. *Let M be a metric space, with distance $d : (\mathbf{x}, \mathbf{x}') \mapsto d(\mathbf{x}, \mathbf{x}')$. Let G be a group acting on M (with, say, a left action). Assume that d is G-invariant, which means that, for all $\mathbf{x}, \mathbf{x}' \in M$ and all $g \in G$,*

$$d(g \cdot \mathbf{x}, g \cdot \mathbf{x}') = d(\mathbf{x}, \mathbf{x}').$$

Then the distance $\bar{d}$ defined on the quotient space M/G by

$$\bar{d}([\mathbf{x}], [\mathbf{x}']) = \inf_{g, g' \in G} d(g \cdot \mathbf{x}, g' \cdot \mathbf{x}') \tag{42}$$

is symmetric and satisfies the triangle inequality.

Notice that, because of the G-invariance, $\bar{d}$ is also given by

$$\bar{d}([\mathbf{x}], [\mathbf{x}']) = \inf_{g \in G} d(g \cdot \mathbf{x}, \mathbf{x}'). \tag{43}$$

A sufficient condition ensuring that $\bar{d}$ is a distance (the missing property being $d([\mathbf{x}], [\mathbf{x}']) = 0 \Rightarrow [\mathbf{x}] = [\mathbf{x}']$) is that the orbits $[\mathbf{x}] = G \cdot \mathbf{x}$ are closed subsets of M for all $\mathbf{x} \in M$. The invariance condition can be placed in parallel with the invariance condition that arose in our discussion of the Riemannian submersion, the latter being an infinitesimal version of the former in the case of Riemannian metrics.

Returning to (41), it is easy to see that a change of offset provides a group action on the left on curves and that the L^2 distance is invariant to this action. It is not too hard to prove that the action has closed orbits so that (41) does provide a valid distance in $\mathcal{B}$. Since the L^2 distance is also invariant by the left action of rotations and translations, one can also define

$$\bar{d}_{L^2}(\mathbf{x},\tilde{\mathbf{x}}) = \inf_{s_0,\theta,b}\left(\int_0^{2\pi} |g_\theta \mathbf{x}(s+s_0)+b-\tilde{\mathbf{x}}(s)|^2 ds\right), \tag{44}$$

where g_θ is the rotation of angle θ and $b \in \mathbb{R}^2$.

A variant of this distance directly compares the derivative of the curves, which provides a translation-invariant representation, defining

$$\bar{d}_{H^1}(\mathbf{x},\tilde{\mathbf{x}}) = \inf_{s_0,\theta}\left(\int_0^{2\pi} |g_\theta \partial_s \mathbf{x}(s+s_0)-\partial_s\tilde{\mathbf{x}}(s)|^2 ds\right). \tag{45}$$

This distance has been introduced for curve comparison in [48], with a very efficient computation algorithm based on Fourier transforms.

When a curve $\mathbf{x}$ is simple (i.e., without self-intersection), it can be considered as the boundary of a bounded set (its interior) that will be denoted $\Omega_{\mathbf{x}}$. A simple distance comparing two such curves, say $\mathbf{x}$ and $\mathbf{x}'$, is the area of the symmetric difference between $\Omega_{\mathbf{x}}$ and $\Omega_{\mathbf{x}'}$, that is,

$$d_{\text{sym}}(\mathbf{x},\mathbf{x}') = \text{area}(\Omega_{\mathbf{x}} \cup \Omega_{\mathbf{x}'}) - \text{area}(\Omega_{\mathbf{x}} \cap \Omega_{\mathbf{x}'}).$$

A more advanced notion, the Hausdorff distance, is defined by

$$d_H(\mathbf{x},\mathbf{x}') = \inf\{\varepsilon > 0, \mathbf{x} \subset B_\varepsilon(\mathbf{x}') \text{ and } \mathbf{x}' \subset B_\varepsilon(\mathbf{x})\},$$

where $B_\varepsilon(\mathbf{x})$ is the set of points at distance less than ε from $\mathbf{x}$ (and similarly for $B_\varepsilon(\mathbf{x}')$). The same distance can be used with $\overline{\Omega}_{\mathbf{x}}$ and $\overline{\Omega}_{\mathbf{x}'}$ instead of $\mathbf{x}$ and $\mathbf{x}'$ for simple closed curves, the Hausdorff distance being in fact a distance between closed subsets of $\mathbb{R}^2$.

Instead of comparing curves that are already parameterized with arc length, one can start with distances that are invariant by reparameterization and quotient out this action as described in Lemma 1. It is not easy to come up with explicit formulae for such invariant distances, but here is an important example.

Start with the supremum norm between the curves, namely,

$$d_\infty(\mathbf{x},\mathbf{x}') = \sup_u |\mathbf{x}(u)-\mathbf{x}'(u)|,$$

which is obviously invariant by changes of parameter. The distance obtained after reduction is called the *Fréchet distance* and is therefore defined by

$$d_F(\mathbf{x},\mathbf{x}') = \inf_\psi d_\infty(\mathbf{x}\circ\psi,\mathbf{x}').$$

Note that, if, for some reparameterization ψ, one has $d_\infty(\mathbf{x}\circ\psi,\mathbf{x}') \le \varepsilon$, then $\mathbf{x} \subset B_\varepsilon(\mathbf{x}')$ and $\mathbf{x}' \subset B_\varepsilon(\mathbf{x})$. This implies the relation

$$\varepsilon > d_F(\mathbf{x},\mathbf{x}') \Rightarrow \varepsilon > d_H(\mathbf{x},\mathbf{x}')$$

which implies $d_H \leq d_F$. As a consequence, $d_F(\mathbf{x}, \mathbf{x}') = 0$ is only possible when $\mathbf{x} = \mathbf{x}'$ up to reparameterization, which completes Lemma 1 in ensuring that d_F is a distance.

Another interesting point of view that leads to parameterization-invariant distances is to include plane curves in a suitable Hilbert space. We have already seen an example of this with the L^2 distance based on the arc-length parameterization, although this one required an extra one-dimensional optimization to get rid of the offset. An interesting alternate option (two of them, in fact) can be obtained by considering curves as linear forms instead of functions.

One can first identify a curve to a measure, which is a linear form on continuous functions, defined by, for a curve, $\mathbf{x}$ and for a function $f : \mathbb{R}^2 \to \mathbb{R}$:

$$(\mu_{\mathbf{x}} | f) = \int_0^{2\pi} f(\mathbf{x}(u)) |\dot{\mathbf{x}}_u(u)| du.$$

This is clearly parameterization independent, and more precisely, $\mu_{\mathbf{x}} = \mu_{\mathbf{y}}$ if and only if $\mathbf{x} = \mathbf{y}$ up to reparameterization or change of orientation.

Another point of view is to identify a curve to a current [19] or, equivalently in this case, to a vector measure which is a linear form on vector fields. For this, simply define, for $f : \mathbb{R}^2 \to \mathbb{R}^2$,

$$(\nu_{\mathbf{x}} | f) = \int_0^{2\pi} \dot{\mathbf{x}}_u(u)^T f(\mathbf{x}(u)) du.$$

This is also parameterization independent, with $\nu_{\mathbf{x}} = \nu_{\mathbf{y}}$ if and only if $\mathbf{x} = \mathbf{y}$ up to reparameterization.

Both (signed) measures and vector measures form linear spaces, even if not all of them correspond to curves. Nonetheless any norm on these spaces directly induces a parameterization-invariant distance between curves. Hilbert norms are specially attracting for this purpose because of the numerical convenience of being associated to a dot product. One way to build such norms is to start with a Hilbert space of functions on $\mathbb{R}^2$ (resp. vector fields) for which $\mu_{\mathbf{x}}$ (resp. $\nu_{\mathbf{x}}$) is continuous and then use the corresponding norm on the dual space [22–25, 73].

Start with the case of scalar functions and consider a Hilbert space W of functions $f : \mathbb{R}^2 \to \mathbb{R}$ such that the evaluation functionals $x \mapsto f(x)$ are continuous (so that W is a reproducing kernel Hilbert space of scalar functions). Denote by $L_W : W \to W^*$ and $K_W : W^* \to W$ the duality operators on W, similarly to what has been introduced in section "A Riemannian Manifold of Deformable Landmarks" with L_V and K_V, so that for $f \in W$ and $\mu \in W^*$,

$$\|f\|_W^2 = (L_W f | f) \text{ and } \|\mu\|_{W^*}^2 = (\mu | K_W \mu).$$

Like in section "A Riemannian Manifold of Deformable Landmarks," K_W is a kernel operator, and there exists a scalar-valued function $(x, y) \mapsto K_W(x, y)$ such that for a measure μ,

$$(K_W \mu)(x) = \int_{\mathbb{R}^2} K_W(x, y) d\mu(y).$$

This implies

$$\|\mu\|^2_{W^*} = \int_{\mathbb{R}^2 \times \mathbb{R}^2} K_W(x, y) d\mu(x) d\mu(y)$$

and directly leads to a distance between curves, namely,

$$\begin{aligned} d(\mathbf{x}, \mathbf{x}')^2 &= \|\mu_{\mathbf{x}} - \mu_{\mathbf{x}'}\|^2_{W^*} \\ &= \int_0^{2\pi} \int_0^{2\pi} K_W(\mathbf{x}(u), \mathbf{x}(u'))|\dot{\mathbf{x}}(u)|\,|\dot{\mathbf{x}}(u')| du\, du' \\ &\quad -2 \int_0^{2\pi} \int_0^{2\pi} K_W(\mathbf{x}(u), \mathbf{x}'(u'))|\dot{\mathbf{x}}(u)|\,|\dot{\mathbf{x}}'(u')| du\, du' \\ &\quad + \int_0^{2\pi} \int_0^{2\pi} K_W(\mathbf{x}'(u), \mathbf{x}'(u'))|\dot{\mathbf{x}}'(u)|\,|\dot{\mathbf{x}}'(u')| du\, du'. \end{aligned} \tag{46}$$

The construction associated to vector measures is similar. The space W being this time a space of vector fields, the discussion is identical to the one holding for V in section "A Riemannian Manifold of Deformable Landmarks," with a kernel K_W which is matrix valued. Other than this, the resulting norm in the dual space is formally the same, yielding

$$\begin{aligned} d(\mathbf{x}, \mathbf{x}')^2 &= \|\nu_{\mathbf{x}} - \nu_{\mathbf{x}'}\|^2_{W^*} \\ &= \int_0^{2\pi} \int_0^{2\pi} \dot{\mathbf{x}}_u^T K_W(\mathbf{x}(u), \mathbf{x}(u'))\dot{\mathbf{x}}_u(u') du\, du' \\ &\quad -2 \int_0^{2\pi} \int_0^{2\pi} \dot{\mathbf{x}}_u^T K_W(\mathbf{x}(u), \mathbf{x}'(u'))\dot{\mathbf{x}}'_u(u') du\, du' \\ &\quad + \int_0^{2\pi} \int_0^{2\pi} \dot{\mathbf{x}}_u'^T K_W(\mathbf{x}'(u), \mathbf{x}'(u'))\dot{\mathbf{x}}'_u(u') du\, du'. \end{aligned} \tag{47}$$

Riemannian Metrics on Curves

We now pass to the specific problem of designing Riemannian metrics on spaces of curves. The first issue we have to deal with is that we are now handling infinite-dimensional manifolds, which is significantly more complex than the finite-dimensional space of landmarks. Since there is more than one type of infinite-dimensional vector spaces, there is more than one type of infinite-dimensional manifolds, and the one which is appropriate when dealing with spaces of infinitely

differentiable curves is the class of *Fréchet manifolds* [31]. It is not our intent, here, to handle the related issues with the appropriate scrutiny, the reader being invited to refer to [54, 55] for a more rigorous presentation. We will here simply state intuitively plausible facts on the structures that are defined.

The space $\mathcal{I}$ of immersed curves is open in the Fréchet space $C^\infty(S^1, \mathbb{R}^2)$ of infinitely differentiable functions from S^1 to $\mathbb{R}^2$ (in which a sequence of curves $\mathbf{x}_n$ converges to $\mathbf{x}$ if all its derivatives converge for the supremum norm). If $\mathbf{x} \in \mathcal{I}$, a tangent vector $\boldsymbol{\xi} \in T_m\mathcal{I}$ is an element of $C^\infty(S^1, \mathbb{R}^2)$ that can also be considered as a smooth vector field along $\mathbf{x}$. A Riemannian metric on $\mathcal{I}$ will therefore be a norm on this space, namely,

$$\boldsymbol{\xi} \mapsto \|\boldsymbol{\xi}\|_{\mathbf{x}}$$

associated to an inner product $\langle \cdot\,, \cdot \rangle_{\mathbf{x}}$ that depends on $\mathbf{x} \in \mathcal{I}$.

We will consider norms that allow for Riemannian projections when quotienting out the action of changes of parameters, as well as the action of the usual transformation groups, $SE(\mathbb{R}^2)$ possibly combined with scaling. Starting with changes of parameters, the differential of the map $\mathbf{x} \mapsto \mathbf{x} \circ \psi$ simply is $\boldsymbol{\xi} \mapsto \boldsymbol{\xi} \circ \psi$, which yields the first requirement

$$\|\boldsymbol{\xi} \circ \psi\|_{\mathbf{x}\circ\psi} = \|\boldsymbol{\xi}\|_{\mathbf{x}} \tag{48}$$

for all $\mathbf{x} \in \mathcal{I}$, $\boldsymbol{\xi} \in C^\infty(S^1, \mathbb{R}^2)$ and smooth reparameterization ψ. A simple way to ensure parameterization invariance is to define the norm for curves that are parameterized with normalized arc length, simply ensuring that the norm is invariant by a change of offset.

Invariance with respect to translations, rotations, and scaling, respectively, requires

$$\|\boldsymbol{\xi}\|_{\mathbf{x}+b} = \|\boldsymbol{\xi}\|_{\mathbf{x}},\ b \in \mathbb{R}^2 \tag{49}$$

$$\|g\boldsymbol{\xi}\|_{g\mathbf{x}} = \|\boldsymbol{\xi}\|_{\mathbf{x}},\ g \in SO(\mathbb{R}^2) \tag{50}$$

$$\lambda\|\boldsymbol{\xi}\|_{\lambda\mathbf{x}} = \|\boldsymbol{\xi}\|_{\mathbf{x}},\ \lambda \in (0, +\infty). \tag{51}$$

A very simple norm, which satisfies (48)–(50), is the L^2 norm of $\boldsymbol{\xi}$ relative to the curve arc length, which is

$$\|\boldsymbol{\xi}\|_{\mathbf{x}}^2 = \int_0^{2\pi} |\boldsymbol{\xi}(u)|^2 |\dot{\mathbf{x}}_u| du. \tag{52}$$

This norm has been studied in [54,55] and shown to provide degenerate Riemannian metrics in the sense that the projected Riemannian distance between any two curves is zero.

Before elaborating on this fact, consider vertical vectors for the projection of $\mathcal{I}$ onto the space $\mathcal{B}$ of curves modulo reparameterization. They are described as

follows. Tangent vectors at $\mathbf{x}$ to the orbit of $\mathbf{x}$ under the action of changes of parameters are obtained as $\boldsymbol{\xi} = (\partial_\varepsilon(\mathbf{x} \circ \psi))(0, u)$, where $\varepsilon \mapsto \psi(\varepsilon, u)$ is a reparameterization in u which smoothly depends on ε. This yields $\boldsymbol{\xi} = \partial_\varepsilon \psi(0, u)\dot{\mathbf{x}}_u \circ \psi(0, u)$, which implies that vertical vectors $\boldsymbol{\xi} \in \mathcal{V}_m$ are such that all $\boldsymbol{\xi}(u)$ are tangent to $\mathbf{x}$.

Horizontal vectors at $\mathbf{x}$ for the metric in (52) are therefore given by vector-valued functions $\boldsymbol{\xi} \mapsto \boldsymbol{\xi}(u)$ that are everywhere normal to $\mathbf{x}$. It follows that if $[\mathbf{x}]$ and $[\mathbf{x}']$ are two equivalent classes of curves modulo reparameterization, their geodesic "distance" is given by

$$d(\mathbf{x}, \mathbf{x}')^2 = \inf \left\{ \int_0^1 \int_0^{2\pi} |\dot{\mathbf{y}}_t|^2 |\dot{\mathbf{y}}_u| du dt, \mathbf{y}(0, .) = \mathbf{x}, [\mathbf{y}(1, .)] = \mathbf{x}', \dot{\mathbf{y}}_u^T \dot{\mathbf{y}}_t = 0 \right\}. \tag{53}$$

As written above, one has the following theorem.

Theorem 1 (Mumford–Michor). *The distance defined in* (53) *vanishes between any pair of smooth curves* $\mathbf{x}$ *and* $\mathbf{x}'$.

A proof of this result can be found in [54, 55]. It relies on the remark that one can grow thin protrusions ("teeth") on a curve at a cost which is negligible compared to the size of the tooth. To get the basic idea underlying this result, one can understand how open segments can be translated at arbitrary small geodesic cost. First, consider a path that starts with a horizontal segment; progressively grow an isosceles triangle of width ε and height t (at time t) somewhere on the segment until $t = 1$. A quick computation shows that the associated geodesic length is $o(\varepsilon)$ (in fact, $O(\varepsilon^2 \ln \varepsilon)$). This implies that one can cover the horizontal segment with $O(1/\varepsilon)$ thin non-overlapping teeth at cost $O(\varepsilon \ln \varepsilon)$. With a similar construction and the same cost, one can pull up the triangles pointing downward to obtain a translated segment. The total cost of the operation being arbitrarily small when $\varepsilon \to 0$, the geodesic distance between parallel segments is zero. This can in fact be extended to any pair of close or open curves, yielding the result stated in Theorem 1.

Quite interestingly, small variations in the definition of the metric are sufficient to address this issue. Take, for example, the distance associated with

$$\|\boldsymbol{\xi}\|_{\mathbf{x}}^2 = \text{length}(\mathbf{x}) \int_0^{2\pi} |\boldsymbol{\xi}(u)|^2 |\dot{\mathbf{x}}_u| du, \tag{54}$$

introduced in [52, 62]. Looking back at the previous "tooth example," the length of a teeth being approximately 2, we see that the length term penalizes the geodesic energy when growing $O(1/\varepsilon)$ teeth by an extra $(1/\varepsilon)$ factor, and the total energy is not negligible anymore. In fact, the associated distance is not degenerate, as shown in [62], in which the geodesic length is proved to correspond to the total area swept by the time-dependent curve.

Another way to control degeneracy is to penalize high curvature points, using, for example,

$$\|\boldsymbol{\xi}\|_{\mathbf{x}}^2 = \int_0^{2\pi} (1 + a\kappa_{\mathbf{x}}(u)^2)|\boldsymbol{\xi}(u)|^2 |\dot{\mathbf{x}}_u| du. \tag{55}$$

This metric has been studied in [55], where it is shown (among other results) that the distance between distinct curves is positive.

All the previous metrics could be put in the form

$$\|\boldsymbol{\xi}\|_{\mathbf{x}}^2 = \int_0^{2\pi} \rho_{\mathbf{x}}(u)|\boldsymbol{\xi}(u)|^2 |\dot{\mathbf{x}}_u| du, \tag{56}$$

where $\rho_{\mathbf{x}} > 0$ is invariant by reparameterization, in the sense that

$$\rho_{\mathbf{x}\circ\psi} \circ \psi = \rho_{\mathbf{x}}.$$

More generally, one can consider metrics associated to positive symmetric linear operators $\boldsymbol{\xi} \mapsto A_{\mathbf{x}}\boldsymbol{\xi}$ which associate to a smooth vector $u \mapsto \boldsymbol{\xi}(u)$ along $\mathbf{x}$ another smooth vector, $u \mapsto (A_{\mathbf{x}}\boldsymbol{\xi})(u)$, with the properties that

$$\int_0^{2\pi} (\boldsymbol{\eta})^T (A_{\mathbf{x}}\boldsymbol{\xi})|\dot{\mathbf{x}}_u| du = \int_0^{2\pi} (A_{\mathbf{x}}\boldsymbol{\eta})^T \boldsymbol{\xi}|\dot{\mathbf{x}}_u| du$$

and

$$A_{\mathbf{x}\circ\psi}(\boldsymbol{\xi} \circ \psi) = (A_{\mathbf{x}}\boldsymbol{\xi}) \circ \psi.$$

The geodesic equation associated to such a metric can be derived by computing the first variation of the geodesic energy. The computation is straightforward if one makes the following assumption on the variations of the operator $A_{\mathbf{x}}$. Assume that there exists a bilinear operator $D'A_{\mathbf{x}}$ that takes as input two vector fields along $\mathbf{x}$, say $\boldsymbol{\xi}(\cdot)$ and $\boldsymbol{\eta}(\cdot)$, and return a new vector $D'A_{\mathbf{x}}(\boldsymbol{\xi}, \boldsymbol{\eta})(\cdot)$ such that

$$\partial_\varepsilon \int_0^{2\pi} (A_{\mathbf{x}+\varepsilon\boldsymbol{\zeta}}\boldsymbol{\xi})^T \boldsymbol{\eta}|\dot{\mathbf{x}}_u| du = \int_0^{2\pi} (D'A_{\mathbf{x}}(\boldsymbol{\xi}, \boldsymbol{\eta}))^T \boldsymbol{\zeta}|\dot{\mathbf{x}}_u| du,$$

where the derivative in the left-hand side is evaluated at $\varepsilon = 0$. With this notation, the geodesic equation is

$$\partial_t(A_{\mathbf{x}}\dot{\mathbf{x}}_t) + (\partial_s\dot{\mathbf{x}}_t)^T \boldsymbol{\tau}\ A_{\mathbf{x}}\dot{\mathbf{x}}_t + \frac{1}{2}\partial_s\Big((A_{\mathbf{x}}\dot{\mathbf{x}}_t)^T \dot{\mathbf{x}}_t\ \boldsymbol{\tau}\Big) = \frac{1}{2}D'A_{\mathbf{x}}(\dot{\mathbf{x}}_t, \dot{\mathbf{x}}_t) \tag{57}$$

with $\partial_s = \partial_u/|\dot{\mathbf{x}}_u|$ as above.

This class of metrics includes the so-called *Sobolev metrics* [52, 56] for which

$$\int_0^{2\pi} (A_{\mathbf{x}}\boldsymbol{\xi})^T \boldsymbol{\xi}\, du = \sum_{k=0}^{p} a_k(\mathbf{x}) \int_0^{2\pi} \left|\partial_s^k \boldsymbol{\xi}\right|^2 du$$

with positive coefficients $a_k(\mathbf{x})$, typically depending on the length of $\mathbf{x}$. Let us take one simple example that has interesting developments: define

$$\int_0^{2\pi} (A_{\mathbf{x}}\boldsymbol{\xi})^T \boldsymbol{\xi}\, du = \text{length}(\mathbf{x})^{-1} \int_0^{2\pi} |\partial_s \boldsymbol{\xi}|^2 du \tag{58}$$

or $A_{\mathbf{x}}\boldsymbol{\xi} = -\text{length}(\mathbf{x})^{-1}\partial_s^2\boldsymbol{\xi}$. The metric associated to $A_{\mathbf{x}}$ is degenerate, since it vanishes over constants. But it provides a metric on curves modulo translations. It satisfies the invariance properties described above, characterized in (48), (50), and (51). This metric was first introduced in [79] and further studied in [68, 69, 81]. A direct computation shows that, in this case,

$$D'A_{\mathbf{x}}(\boldsymbol{\xi}, \boldsymbol{\eta}) = 2\text{length}(\mathbf{x})^{-1}\partial_s((\partial_s\boldsymbol{\xi})^T\partial_s\boldsymbol{\eta}\boldsymbol{\tau}) - \text{length}(\mathbf{x})^{-1}\langle\boldsymbol{\xi}\,,\,\boldsymbol{\eta}\rangle_{\mathbf{x}}\boldsymbol{\kappa}\boldsymbol{\nu}.$$

The study of this metric is, however, much simpler than replacing the expression of $D'A_{\mathbf{x}}$ into (57) would make believe. The simplification comes after the following transformation of the curve representation. Consider the transformation, defined over pairs of real-valued functions $u \mapsto (\mathbf{a}(u), \mathbf{b}(u))$ by

$$\mathbf{x}(u) = \left(\frac{1}{2}\int_0^u (\mathbf{a}^2 - \mathbf{b}^2) d\tilde{u}\,,\, \int_0^u \mathbf{a}\mathbf{b}\, d\tilde{u}\right),\ u \in [0, 2\pi], \tag{59}$$

so that

$$\dot{\mathbf{x}}_u = ((a^2 - b^2)/2\,,\, ab).$$

With the notation above, we have $|\dot{x}_u| = (a^2 + b^2)/2$. This generate a curve in $\mathbb{R}^2$, with length

$$\text{length}(\mathbf{x}) = \frac{1}{2}\int_0^{2\pi} (\mathbf{a}^2 + \mathbf{b}^2) du = \frac{1}{2}\left(\|\mathbf{a}\|_2^2 + \|\mathbf{b}\|_2^2\right).$$

Denoting by $\mathbf{x} = T(\mathbf{a}, \mathbf{b})$ the transformation in Eq. (59), one can write the differential of T as

$$DT(\mathbf{a}, \mathbf{b})(\boldsymbol{\alpha}, \boldsymbol{\beta}) : u \mapsto \left(\int_0^u (\mathbf{a}\boldsymbol{\alpha} - \mathbf{b}\boldsymbol{\beta}) d\tilde{u}\,,\, \int_0^u (\mathbf{b}\boldsymbol{\alpha} + \mathbf{a}\boldsymbol{\beta}) d\tilde{u}\right)$$

and a direct computation yields

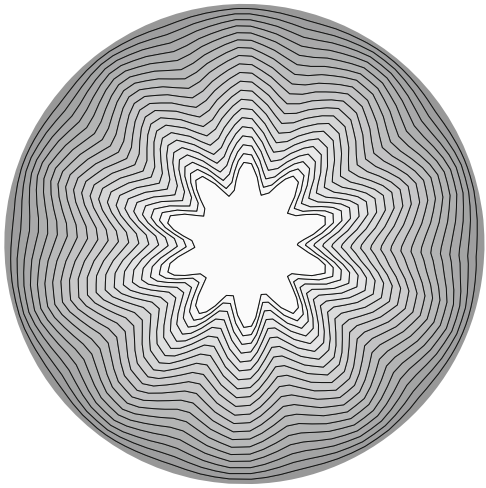


Fig. 4 An example of a geodesic connecting a circle to a star-shaped curve for the metric defined in (58). The evolving curves are superimposed with progressively reduced size to facilitate visualization (the compared curves having both length 1 originally)

$$\|DT(\mathbf{a},\mathbf{b})(\boldsymbol{\alpha},\boldsymbol{\beta})\|_{T(\mathbf{a},\mathbf{b})} = 2\frac{\sqrt{\|\boldsymbol{\alpha}\|_2^2 + \|\boldsymbol{\beta}\|_2^2}}{\sqrt{\|\mathbf{a}\|_2^2 + \|\mathbf{b}\|_2^2}}.$$

Restricting to closed curves with unit length implies the conditions

$$\|\mathbf{a}\|_2 = \|\mathbf{b}\|_2 = 1 \text{ and } \langle \mathbf{a}\,,\,\mathbf{b}\rangle_2 = 0$$

which means that $(\mathbf{a},\mathbf{b})$ forms an orthonormal two-frame in the space $L^2(S^1)$, that is, an element of the Stiefel manifold $\mathrm{St}(L^2,2)$. Up to the factor two, the mapping T is then an isometry between $\mathrm{St}(L^2,2)$, equipped with its standard metric, and the subset of $\mathcal{I}$ consisting of unit-length curves. If one furthermore makes the reduction of quotienting out rotations for curves, one finds that the isometry becomes with the Grassmannian manifold $\mathrm{Gr}(L^2,2)$ of two-dimensional subspaces of L^2. This identification can be exploited to obtain explicit geodesics in the considered shape space (see [81]). It is important to notice that the restriction to curves with unit length is equivalent to making the Riemannian projection on the quotient space modulo scalings. This is because horizontal vectors for the scale action can easily be shown to satisfy $\int (\partial_s \xi)^T \tau = 0$, which, if $\xi = \dot{\mathbf{x}}_t$, directly implies that $\partial_t(\int |\dot{\mathbf{x}}_u|^2) = 0$. Therefore, length is conserved along horizontal geodesics, which justifies the choice of unit-length curves. Some numerical issues associated to this metric are studied in [68] and [69] in the simpler case in which it is applied to open curves. An example of geodesic obtained using this metric is provided in Fig. 4.

A parameterized variant of this metric, applied to closed curves with unit length, has been proposed in [43], in the form

$$\|\boldsymbol{\xi}\|_{\mathbf{x}}^2 = \int_0^{2\pi} ((\partial_s)\boldsymbol{\xi}^T\tau)^2|\dot{\mathbf{x}}_u|du + c\int_0^{2\pi} ((\partial_s)\boldsymbol{\xi}^T\nu)^2|\dot{\mathbf{x}}_u|du,$$

the previous metric corresponding to $c = 1$. When $c \neq 1$, the unit length constraint is not induced by a Riemannian projection, but the metric can be studied on this space anyway.

One can analyze this metric in the following way. Let $\mathbf{x}(t,u)$ be a time-dependent curve. Define $\lambda = |\dot{\mathbf{x}}_u|$ so that

$$\partial_s\dot{\mathbf{x}}_t = \lambda^{-1}\partial_t(\lambda\tau) = \partial_t(\log\lambda)\tau + \partial_t\tau.$$

Since the two terms in the sum are perpendicular, this gives

$$\|\dot{\mathbf{x}}_t\|_{\mathbf{x}}^2 = \int_0^{2\pi} \left((\partial_t\log\lambda)^2 + c|\partial_t\tau|^2\right)|\dot{\mathbf{x}}_u|du. \tag{60}$$

The first term measures the logarithmic variation of the arc length, and the second is the instantaneous rotation of the tangents. Interestingly, another change of variable akin to the one discussed for $c = 1$ can also simplify this metric in the case $c = 4$. Take, in this case,

$$\mathbf{x} = T(a,b) := u \mapsto \left(\int_0^u a\sqrt{a^2+b^2}du',\ \int_0^u b\sqrt{a^2+b^2}du'\right);$$

one has this time

$$\|DT(a,b)(\alpha,\beta)\|_{T(a,b)}^2 = 4\int_0^{2\pi}(\alpha^2+\beta^2)du$$

which provides an identification of the space of open curves with unit length with an infinite-dimensional sphere. This identification has the important property to carry over to higher-dimensional curves [37]. There is, however, no "nice" representation for closed curves in this case.

Notice that the two identifications that were just discussed apply to parameterized curves. In both cases, the geodesic distance must be optimized with respect to reparameterization to obtain a metric between geometric curves.

Another important contribution to the theory of spaces of plane curves was made in [63], in which simple closed domains in $\mathbb{R}^2$ are represented via the correspondence maps between the conformal mapping of their interior and of their exterior to the unit disc. This induces an almost one-to-one representation of simple curves by diffeomorphisms of the unit circle. In fact, this representation has to come modulo Möbius transformations on the circle, which are very simply accommodated by an invariant metric, called the Weil–Peterson metric, on such diffeomorphisms. The reader is referred to the cited work for more details.

Projecting the Action of 2D Diffeomorphisms

At the exception of the one just mentioned, the previously discussed metrics were all defined based on the parameterizations of the curves. This provided reasonably simple definitions, exploiting in particular the invariance property of the arc length. Because they relied on local properties of the curves, these metrics were not able to penalize singularities that occur globally, like the intersection of two remote parts.

One way to handle global constraints is to use an approach similar to the one that has been used to define the landmark manifold, based on the action of two-dimensional diffeomorphisms on curves. This will therefore be based on the projection paradigm discussed in section "Reduction: Transitive Group Action."

So, let $G \subset \mathrm{Diff}(\mathbb{R}^2)$ be a group of smooth diffeomorphisms of $\mathbb{R}^2$ (which, say, smoothly converge to the identity at infinity), and let $\mathbf{x}_0$ be a reference curve or template. Consider the set $M = G \cdot \mathbf{x}_0$, the orbit of $\mathbf{x}_0$ under the action of G, the latter being simply defined by

$$(\varphi \cdot \mathbf{x})(u) = \varphi(\mathbf{x}(u)).$$

This implies that $D\pi(\varphi)v = v \circ \mathbf{x}_0$ and a horizontal covector at $\varphi \in G$ for the projection takes the form

$$(p|\, v) = (\rho|\, v \circ \mathbf{x}_0)$$

for some $\rho \in T_{\varphi(\mathbf{x}_0)} M^*$.

Let's make this explicit for ρ belonging to an important class of linear forms on $T_{\mathbf{x}} M$, associated to vector measures, that is,

$$(\rho|\, \xi) = \int_0^{2\pi} \xi^T \mathbf{a} d\mu$$

where μ is a measure on the unit circle and $\mathbf{a}$ is a vector-valued function. The associated horizontal covector is then

$$(p|\, v) = \int_0^{2\pi} v(\mathbf{x}_0(u))^T \mathbf{a}(u) d\mu(u) \tag{61}$$

and the reduced Hamiltonian computed on this covector is (denoting as in section "Reduction: Transitive Group Action" K_φ the duality operator on $T_\varphi G$, still assumed to be associated to a reproducing kernel)

$$H_M(\mathbf{x}, \rho) = \frac{1}{2} \int_0^{2\pi} \int_0^{2\pi} \mathbf{a}(u)^T K_\varphi(\mathbf{x}_0(u), \mathbf{x}_0(u')) \mathbf{a}(u') d\mu(u') d\mu(u) \tag{62}$$

with $\mathbf{x} = \varphi \cdot \mathbf{x}_0$. As in section "Reduction: Transitive Group Action," the invariance requirement boils down to $K_\varphi(\mathbf{x}_0(u), \mathbf{x}_0(u'))$ only depending on $\varphi \cdot \mathbf{x}_0$, with the simplest choice associated to a right-invariant metric on G, yielding

$$H_M(\mathbf{x},\rho)=\frac{1}{2}\int_0^{2\pi}\int_0^{2\pi}\mathbf{a}(u)^T K_V(\mathbf{x}(u),\mathbf{x}(u'))\mathbf{a}(u')d\mu(u')d\mu(u) \tag{63}$$

for a fixed kernel K_V. An important fact is that measure covectors remain so during the evolution, the Hamiltonian (or geodesic) equations are simply written as

$$\begin{cases}\partial_t\mathbf{x}(t,u)=\displaystyle\int_0^{2\pi}K_V(\mathbf{x}(t,u),\mathbf{x}(t,\tilde{u}))\mathbf{a}(t,\tilde{u})d\mu(\tilde{u})\\ \partial_t\mathbf{a}(t,u)=-\displaystyle\sum_{i,j=1}^{2}\int_0^{2\pi}\mathbf{a}^i(t,u)\mathbf{a}^j(t,\tilde{u})\nabla_1 K^{ij}(\mathbf{x}(t,u),\mathbf{x}(t,\tilde{u}))d\mu(\tilde{u})\end{cases} \tag{64}$$

Another interesting fact is that (64) exactly provides (6) in the case when μ is a weighted sum of Dirac measures. This is because these equations are, as proved in section "Reduction via a Submersion," all particular instances of the Hamiltonian system (or geodesic equation) obtained on the acting group of diffeomorphisms, namely, (25).

This was the first step downward, from diffeomorphisms to parameterized plane curves. It remains to discuss the additional steps, which are the reduction for the required invariance, by reparameterization and Euclidean transformation.

Consider the action of reparameterization, which is a right action. The action of change of parameters on vector measures like in (61) is

$$(p\cdot\psi\,|\,\xi)=\int_0^{2\pi}\mathbf{a}^T\xi\circ\psi^{-1}d\mu(u)=\int_0^{2\pi}(a\circ\psi)^T\xi d(\psi^{-1}\mu)(u),$$

where $\psi\cdot\mu$ is the image of μ by ψ. Using this, the invariance requirement applied to a Hamiltonian taking the form

$$H_M(\mathbf{x},\rho)=\frac{1}{2}\int_0^{2\pi}\int_0^{2\pi}\mathbf{a}(u)^T K_{\mathbf{x}}(u,u')\mathbf{a}(u')d\mu(u')d\mu(u) \tag{65}$$

can be seen to reduce to the constraint that

$$K_{\mathbf{x}\circ\psi}(\psi^{-1}(u),\psi^{-1}(u))=K_{\mathbf{x}}(u,u')$$

and this property is satisfied for $K_{\mathbf{x}}(u,u)=K_V(\mathbf{x}(u),\mathbf{x}(u'))$.

The momentum map associated to changes of parameters is

$$(\mathfrak{m}(\mathbf{x},p)\,|\,v)=\int_0^{2\pi}\mathbf{a}(u)^T\dot{\mathbf{x}}_u v(u)d\mu(u),$$

so that horizontal vector measures simply are those for which $\mathbf{a}$ is normal to the curve, that is, $\mathbf{a}(u)=\boldsymbol{\alpha}(u)\boldsymbol{\nu}(u)$, where $\boldsymbol{\alpha}$ is scalar valued and $\boldsymbol{\nu}$ is the normal to $\mathbf{x}$. The evolution equations then become

$$\begin{cases} \partial_t \mathbf{x}(t,u) = \displaystyle\int_0^{2\pi} K_V(\mathbf{x}(t,u),\mathbf{x}(t,\tilde{u}))\boldsymbol{\alpha}(t,\tilde{u})\boldsymbol{\nu}(t,\tilde{u})d\tilde{u} \\ \partial_t \boldsymbol{\alpha}(t,u) = -\boldsymbol{\alpha}(t,u)\displaystyle\sum_{i,j=1}^{2}\int_0^{2\pi} \boldsymbol{\alpha}(t,\tilde{u})\boldsymbol{\nu}^i(t,u)\boldsymbol{\nu}^j(t,\tilde{u})\nabla_1 K^{ij}(\mathbf{x}(t,u),\mathbf{x}(t,\tilde{u}))d\tilde{u}. \end{cases} \tag{66}$$

If K_V is furthermore invariant by rotation and translation, quotienting out these operations results in additional conditions on $\boldsymbol{\alpha}$. Invariance by translation requires

$$\int_0^{2\pi} \boldsymbol{\alpha}(u)\boldsymbol{\nu}(u)du = 0,$$

and the constraint associated to rotations is

$$\int_0^{2\pi} \boldsymbol{\alpha}(u)(\boldsymbol{\nu}(u)\mathbf{x}(u)^T - \mathbf{x}(u)\boldsymbol{\nu}(u)^T)du = 0.$$

Extension to More General Shape Spaces

The construction based on the Riemannian submersion from groups of diffeomorphisms can be reproduced in a large variety of contexts, essentially for any class of objects that can be deformed by diffeomorphisms. This can be applied to provide metrics on space of surfaces and spaces of images, of vector fields, of measures, etc.

Let us consider, for example, the case of images that we will take as differentiable functions $I : \mathbb{R}^d \to \mathbb{R}$. Define the left action of a diffeomorphism φ on an image I to be

$$\varphi \cdot I = I \circ \varphi^{-1}.$$

From this, one sees that the infinitesimal action of a vector field V on I is

$$v \cdot I = -v^T \nabla I.$$

(This is why we assumed that the images are differentiable. For non-differentiable images, $v \cdot I$ is not a function, but a distribution, with, if ρ is a smooth function,

$$(v \cdot I \,|\, \rho) = \int_{\mathbb{R}^d} I \nabla \cdot (\rho v)dx,$$

where $\nabla \cdot$ is the divergence operator. The reader is referred to [75, 76] for the analysis of the inexact matching approach in the more general case of images with bounded variations.)

Fix a reference image I_0 and consider the space

$$M = \{\varphi \cdot I_0, \varphi \in G\},$$

the surjection being as usual $\pi(\varphi) = \varphi \cdot I_0$. Consider covectors on M that are associated to measures, namely,

$$(\rho|\,\xi) = \int_{\mathbb{R}^d} \xi(x) d\rho(x),$$

where ξ is a real-valued function (which represents a tangent vector to M). The differential of $\pi(\varphi) = I_0 \circ \varphi^{-1}$ is (letting $\psi = \varphi^{-1}$)

$$D\pi(\varphi)v = -(\nabla I_0 \circ \varphi^{-1}) D(\varphi^{-1}) v \circ \varphi^{-1} = -\nabla I^T v \circ \varphi^{-1}$$

with $I = \varphi \cdot I_0$, so that the horizontal covector at $\varphi \in G$ associated to a measure ρ is $p = D\pi(\varphi)^* \rho$ defined by

$$(p|\,v) = -\int_{\mathbb{R}^d} v(\varphi^{-1}(x))^T \nabla I(x) d\rho(x),$$

where $I = \varphi \cdot I_0$. Starting from a Hamiltonian associated to a right-invariant metric on G yields the reduced Hamiltonian

$$\begin{aligned} H_M(I,\rho) &= \frac{1}{2} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \nabla I(x)^T K_\varphi(\varphi^{-1}(x), \varphi^{-1}(y)) \nabla I(y) d\rho(y) d\rho(x) \\ &= \frac{1}{2} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \nabla I(x)^T K_V(x,y) \nabla I(y) d\rho(y) d\rho(x) \end{aligned}$$

with $K_\varphi(x,y) = K_V(\varphi(x), \varphi(y))$. The associated Hamiltonian equations are

$$\begin{cases} \partial_t I(x) = \int_{\mathbb{R}^d} \nabla I(x)^T K_V(x,y) \nabla I(y) d\rho(y) dy \\ \partial_t \alpha = \nabla \cdot (\alpha K_V(\nabla I \rho)) \end{cases}$$

A limitation in the image case is that two given images are very rarely connected by diffeomorphisms, so that working with images that are deformations of a reference image is a strong restriction. This issue can be addressed by extending the projection to a larger set than the sole group of diffeomorphisms. One can use a simple construction for this: call M the space of *all* smooth images (instead of just an orbit, as it was defined before), and still let G denote a group of smooth diffeomorphisms. Consider the surjection $\pi : G \times M \to M$ defined by

$$\pi(\varphi, I) = \varphi \cdot I.$$

(This is obviously a surjection since $I = \pi(\mathrm{id}_{\mathbb{R}^d}, I)$.)

Letting $J = I \circ \varphi^{-1}$, one has

$$D\pi(\varphi, I)(v, \xi) = -\nabla J^T v \circ \varphi^{-1} + \xi \circ \varphi^{-1},$$

so that the horizontal covector at (φ, I) associated to a measure ρ on M is $\bar{p} = D\pi(\varphi, I)^*\rho$ such that

$$(\bar{p}|\,(v,\xi)) = \int_{\mathbb{R}^d} (-\nabla J^T v \circ \varphi^{-1} + \xi \circ \varphi^{-1}) d\rho$$

with $J = I \circ \varphi^{-1}$. Thinking of a covector $\bar{p} \in T_{(\varphi,I)}(G \times M)^*$ as a pair (p, η) with $p \in T_\varphi G^*$ and $\eta \in T_I M^*$, one can identify $(p, \eta) = D\pi(\varphi, I)^*\rho$ as

$$(p|\,v) = -\int_{\mathbb{R}^d} \nabla J^T v \circ \varphi^{-1} d\rho \text{ and } (\eta|\,\xi) = \int_{\mathbb{R}^d} \xi \circ \varphi^{-1} d\rho.$$

If $d\rho = \rho dx$ is absolutely continuous with respect to Lebesgue's measure, the second term gives $\eta = \rho \circ \varphi \det D\varphi\, dx$.

If one starts with a Hamiltonian on $G \times M$ for which

$$H((\varphi, I), (p, \eta dx)) = \frac{1}{2}\big(p|\,K_\varphi p\big) + \frac{\lambda}{2}\int_{\mathbb{R}^d} \eta(x)^2 (\det D\varphi(x))^{-1} dx$$

with K_φ as above, the resulting reduced Hamiltonian on M is

$$H_M(J, \rho dx) = \frac{1}{2}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \rho(x)\nabla J(x)^T K_V(x,y)\nabla J(y)\rho(y) dx dy + \frac{\lambda}{2}\int_{\mathbb{R}^d} \rho(x)^2 dx.$$

The corresponding evolution equations then are

$$\begin{cases} \partial_t J = \nabla J^T K_V(\rho \nabla J) + \lambda\rho \\ \partial_t \rho = \nabla \cdot (\rho K_V(\rho \nabla J)). \end{cases}$$

This is a particular instance of the theory of metamorphosis applied to images (the interested reader can refer to [35, 71] for further developments).

Applications to Statistics on Shape Spaces

An important situation in which the previously discussed concepts are relevant is for the analysis of shape samples, that is, families $\mathbf{x}_1, \ldots, \mathbf{x}_n$, in which each $\mathbf{x}_j$ is a shape, possibly represented as a collection of landmarks or a plane curve (or another representation, like surfaces, images, etc.), and interpreted as a point in a manifold M. A simple and commonly used approach to analyze such samples is to "normalize" them using the exponential or momentum representation relative to a fixed template $\bar{\mathbf{x}}$. Each shape $\mathbf{x}_k$ is then transformed into a tangent or cotangent vector, say $\boldsymbol{\xi}_k \in T_{\bar{\mathbf{x}}} M$ so that $\mathbf{x}_k = \exp_{\bar{\mathbf{x}}}(\boldsymbol{\xi}_k)$.

The problem is then reduced to the well-explored context of data analysis in a linear space, and how it is analyzed afterward depends on the specific problem at hand and is out of the scope of the present discussion. An important thing that one should remember is that this reduction can be accompanied with significant metric distortion, related to curvature as described in section "Metric Distortion and Curvature" (in spaces with positive curvature, the representation may even fail to be one to one). The approach has however proved to be a powerful analysis tool in several applications [20,42], including the analysis of medical data [74].

This distortion being larger when the distances between the represented shapes and the template are large, it is natural to select the template in a way that minimizes these distances, the most widespread approach being to define it as a Karcher (or geometric) mean, that is, as a minimizer of

$$U(\mathbf{x}) = \sum_{k=1}^{n} d_M(\mathbf{x}, \mathbf{x}_k)^2. \tag{67}$$

This well-posedness of this definition is also related to the curvature. The function U is convex and the minimum is unique if M has negative curvature [39]. Negative curvature is unfortunately difficult to obtain in shape spaces because the reduction process always increases the sectional curvature [59] (notice however that the representation in [63] has negative curvature, but it seems to be the only such example). The sectional curvature on the landmark manifold, as shown in [53], can be both positive and negative. As proved in [39], a sufficient condition ensuring the convexity of U (67) is that the diameter of the sample set (the largest geodesic distance between two of the points) is smaller than $\pi/\left(2\sqrt{s_{\max}}\right)$, where $s_{\max}$ is a positive upper bound of the sectional curvature (U is always convex with negative curvature). Interestingly, in that case, the optimality condition of the Karcher mean is that it constitutes a sample average in the exponential representation, that is, $\mathbf{x}_k = \exp_{\bar{\mathbf{x}}}(\boldsymbol{\xi}_k)$ with $\sum_{k=1}^{n} \boldsymbol{\xi}_k = 0$. This leads to an algorithm for the computation of the mean, which can be proved to converge under similar curvature conditions [46,47]: start with an initial guess for $\bar{\mathbf{x}}$ and compute the exponential representation $\boldsymbol{\xi}_k$ over the sample set. Compute $\bar{\boldsymbol{\xi}} = \sum_{k=1}^{n} \boldsymbol{\xi}_k/n$, replace $\bar{\mathbf{x}}$ by $\exp_{\bar{\mathbf{x}}}(-\bar{\boldsymbol{\xi}})$, and iterate until stabilization. A variation of this algorithm has been proposed in [21]. One can also mention the interesting algorithm proposed in [13] in which kernel regression is generalized to shape manifolds.

4 Numerical Methods and Case Examples

The most important numerical method on the previously discussed shape spaces is related to the computation of geodesics (i.e., solving the geodesic equation) and, most importantly in practice, to the computation of the representation in the exponential chart or of the momentum representation.

This section will focus on the latter problem (which anyway includes the first one as a subproblem) that will first be addressed in the simpler case of the landmark manifold.

To compute exponential coordinates around some object $\mathbf{x}_0$, one needs to solve, for some target object $\mathbf{y}$, the equation

$$\exp_{\mathbf{x}_0}(\boldsymbol{\xi}) = \mathbf{y} \tag{68}$$

or, if the momentum representation is more convenient,

$$\exp^{\flat}_{\mathbf{x}_0}(\boldsymbol{\alpha}) = \mathbf{y}. \tag{69}$$

Since these representations are defined by nonlinear evolution equations, this is a highly nonlinear problem, in which the function to be inverted cannot be written in closed form. Also, in the case of curves, the problem is infinite dimensional and must therefore be properly discretized. Another non-negligible issue is that, even if the equation has a solution (which is often the case in the discussed framework), this solution is not necessarily unique unless $\mathbf{y}$ is close enough to $\mathbf{x}_0$. For this reason, it may be impossible to represent a generic shape dataset using only one of these charts, but this may be achievable for a more focused one (like shapes of fish, or leaves, of fixed anatomical organs).

There are mainly two options to address the computation. The first one is to directly solve the equation (using zero-finding methods, like Newton's algorithm). The second one is to return to the definition of geodesics as curves with minimal energy and to solve the variational problem of finding minimal energy paths between $\mathbf{x}_0$ and $\mathbf{y}$.

Landmark Matching via Shooting

Let us start with the first approach. Recall that given some differentiable function $F : \mathbb{R}^n \to \mathbb{R}^n$, Newton's method to solve the equation $F(z) = 0$ iterates (starting with a good guess of the solution, z_0)

$$z_{k+1} = z_k - DF(z_k)^{-1} F(z_k).$$

This scheme can be directly applied to the solution of (69) in the landmark case, with $F(\boldsymbol{\alpha}_0) = \exp^{\flat}_{\mathbf{x}_0}(\boldsymbol{\alpha}_0) - \mathbf{y}$, since it is finite dimensional; one needs to compute the differential of the momentum representation, which is only described in (6) via the solution of a differential equation. As a result, the differential of F, which is also the differential of $\exp^{\flat}_{x_0}$, must also be computed by solving a differential equation. Noting that (6) takes the form

$$\begin{cases} \partial_t \mathbf{x} = Q(\mathbf{x}, \boldsymbol{\alpha}) \\ \partial_t \boldsymbol{\alpha} = R(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\alpha}) \end{cases} \tag{70}$$

with Q linear and R quadratic in $\boldsymbol{\alpha}$, and that $\mathbf{x}(t) = \exp^{\flat}_{\mathbf{x}_0}(t\boldsymbol{\alpha}_0)$, we have denoting

$$J(t) = D\exp^{\flat}_{\mathbf{x}_0}(t\boldsymbol{\alpha}_0) \tag{71}$$

$$\begin{cases} \partial_t J\boldsymbol{\beta} = \partial_1 Q(\mathbf{x}, \boldsymbol{\alpha}) J\boldsymbol{\beta} + Q(\mathbf{x}, H\boldsymbol{\beta}) \\ \partial_t H\boldsymbol{\beta} = \partial_1 R(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\alpha}) J\beta + 2R(\mathbf{x}, \boldsymbol{\alpha}, H\boldsymbol{\beta}) \end{cases} \tag{72}$$

in which H is an auxiliary operator that represents the variation in $\boldsymbol{\alpha}$ (and ∂_1 is the differential with respect to the first variable). Solving (70) and (72) up to time $t = 1$ provides $\mathbf{x}(1)$ and $J(1)$, and the Newton step is given by

$$\boldsymbol{\alpha}_0^{k+1} = \boldsymbol{\alpha}_0^k - J(1)^{-1}(\mathbf{x}(1) - \mathbf{y}).$$

Making explicit the expressions of $\partial_1 Q$ and $\partial_1 R$ is not difficult, but rather lengthy, and these expressions will not be provided here (the interested reader can refer to [1] for more details). An important limitation for the feasibility of this kind of approach is the cost involved in the computation of the full matrix $J(t)$. With N landmarks in d dimensions, the size of $\mathbf{x}$ and $\boldsymbol{\alpha}$ is $n = Nd$ and the size of J is n^2. The computation of the right-hand size of (72) requires an order of n^3 operations if one takes advantages of the special structure of the operator $Q(\mathbf{x}, \cdot)$ (it would be n^4 otherwise). Even with this reduction, a computation cost which is cubic in the number of landmarks rapidly becomes unfeasible, and it is difficult to run this algorithm with, say, more than a few hundred landmarks. On the other hand, convergence (when it happens) can require a very small number of steps.

Another limitation of Newton's method is the fact that it is not guaranteed to converge, unless the starting point ($\boldsymbol{\alpha}_0^0$ with our notation) is close enough to the solution, in a way which is generally impossible to quantify a priori. For this reason, the method is often usefully complemented (and possibly replaced if the number of landmarks is too large) by simple gradient descent in which the minimized function is

$$F(\boldsymbol{\alpha}_0) = \left(\exp^{\flat}_{\mathbf{x}_0}(\boldsymbol{\alpha}_0) - \mathbf{y}\right)^T \left(\exp^{\flat}_{\mathbf{x}_0}(\boldsymbol{\alpha}_0) - \mathbf{y}\right).$$

The first variation of F is, with the previous notation,

$$\partial_\varepsilon F(\boldsymbol{\alpha}_0 + \varepsilon\boldsymbol{\beta})_{|[\varepsilon=0]} = 2\left(\exp^{\flat}_{\mathbf{x}_0}(\boldsymbol{\alpha}_0) - \mathbf{y}\right)^T J(1)\boldsymbol{\beta}.$$

It is natural to define gradients relative to the Riemannian metric at $\mathbf{x}_0$, as defined in Eq. (4). When working with momenta as done here, the gradient should be identified using

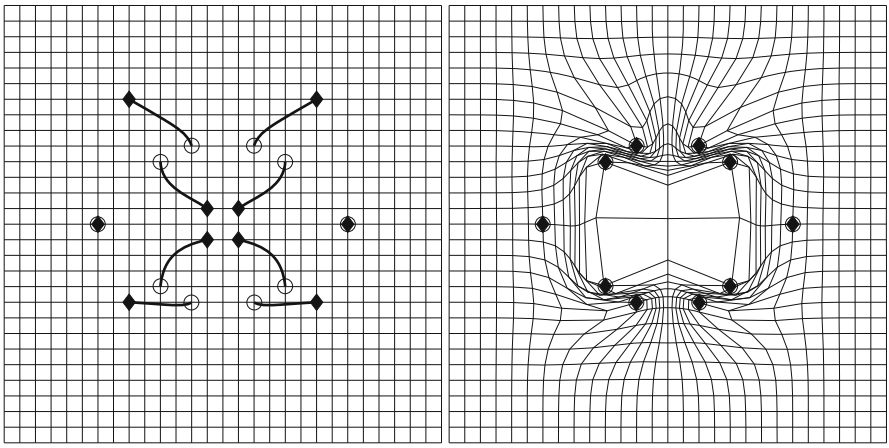


Fig. 5 Example of geodesics between two landmark configurations: *left*, trajectories (diamonds move onto circles), and *right*, resulting diffeomorphism

$$\boldsymbol{\beta}^T S_V(\mathbf{x}_0)\nabla F(\boldsymbol{\alpha}_0) = 2\left(\exp^{\flat}_{\mathbf{x}_0}(\boldsymbol{\alpha}_0) - \mathbf{y}\right)^T J(1)\beta$$

yielding

$$\nabla F(\boldsymbol{\alpha}_0) = 2S_V(\mathbf{x}_0)^{-1}\left(J(1)^T\left(\exp^{\flat}_{\mathbf{x}_0}(\boldsymbol{\alpha}_0) - \mathbf{y}\right)\right). \tag{73}$$

The computation, for a given vector $\mathbf{z}$, of $J(t)^T\mathbf{z}$ can be done by solving backward in time the system

$$\begin{cases} \partial_t \boldsymbol{\xi} = -(\partial_1 Q(\mathbf{x},\alpha))^T\boldsymbol{\xi} - (\partial_1 R(\mathbf{x},\alpha,\alpha))^T\mathbf{a} \\ \partial_t \mathbf{a} = -Q(\mathbf{x})^T\boldsymbol{\xi} - 2R(\mathbf{x},\alpha)^T\mathbf{a} \end{cases} \tag{74}$$

initialized with $(\boldsymbol{\xi}(1),\mathbf{a}(1)) = (\mathbf{z},0)$, with the notation $Q(\mathbf{x})\boldsymbol{\beta} = Q(\mathbf{x},\boldsymbol{\beta})$ and $R(\mathbf{x},\boldsymbol{\alpha})\boldsymbol{\beta} = R(\mathbf{x},\boldsymbol{\alpha},\boldsymbol{\beta})$. One then has $J(1)^T\mathbf{z} = \mathbf{a}(0)$. The proof of this statement derives from elementary computations on linear dynamical systems.

This implies that the term $J(1)^T\left(\exp^{\flat}_{\mathbf{x}_0}(\boldsymbol{\alpha}_0) - \mathbf{y}\right)$ can be computed by solving an ODE which has the same dimension as the geodesic Eq. (70). Notice, however, that (74) requires using the solution of (70) with a backward time evolution (from $t = 1$ to $t = 0$). This implies that the solution of (70) must be first computed and stored with a fine enough time discretization to allow for an accurate solution of (74). This may cause memory issues for high-dimensional models. An example of trajectories and deformations estimated using this algorithm is provided in Fig. 5.

The above discussion only addressed the computation of geodesics in landmark shape space without quotienting out rotations and translations. Recall that this operation, when done starting from a metric for which the projection on the quotient space is a Riemannian submersion, only requires to constrain the momentum

representation with a finite number of linear relations. The associated reduction in the number of degrees of freedom is balanced by the reduced requirement of connecting the reference shape to some element of the orbit of the target under the quotiented out group action, instead of the target itself. More explicitly, the equations that need to be solved to compute the momentum representation of $\mathbf{y}$ relative to $\mathbf{x}_0$ are

$$\begin{cases} \exp^{\flat}_{\mathbf{x}_0}(\boldsymbol{\alpha}_0) - g \cdot \mathbf{y} = 0 \\ \sum_{k=1}^{N} \alpha_{0,k} = 0 \\ \sum_{k=1}^{N} \left(\alpha_{0,k} x_{0,k}^T - x_{0,k} \alpha_{0,k}^T\right) = 0, \end{cases} \tag{75}$$

where $g \in SE(\mathbb{R}^d)$. A transformation g in this space is represented by a rotation part, R, and a translation part, b, and classically parameterized in the form

$$\begin{pmatrix} R & b \\ 0 \cdots 0 & 1 \end{pmatrix} = \exp \begin{pmatrix} A & \omega \\ 0 \cdots 0 & 0 \end{pmatrix}$$

with A skew symmetric and $\omega \in \mathbb{R}^d$. System (75) therefore has $Nd + d(d+1)/2$ equations and variables and can be solved as above using Newton iterations when feasible or gradient descent. Since the exponential is the solution of a differential equation $\left(\partial_t \exp(tU) = U \exp(tU)\right)$, optimization in A and ω above can be treated exactly like the optimization in $\boldsymbol{\alpha}_0$. Another option is to directly use the formula

$$\partial_\varepsilon \exp(U + \varepsilon h)_{|_{\varepsilon=0}} = \int_0^1 \exp(tU) h \exp(-tU) dt.$$

Landmark Matching via Path Optimization

The other option, in order to compute the momentum representation, is to solve the shortest path problem between $\mathbf{x}_0$ and $\mathbf{y}$, that is, to minimize

$$E(\mathbf{x}(\cdot)) = \int_0^1 \|\dot{\mathbf{x}}_t\|^2_{\mathbf{x}(t)} dt,$$

with the constraints $\mathbf{x}(0) = \mathbf{x}_0$ and $\mathbf{x}(1) = \mathbf{y}$, using gradient descent on the space of all trajectories $t \mapsto \mathbf{x}(t)$. Letting $P(\mathbf{x}, \boldsymbol{\xi}) = \|\boldsymbol{\xi}\|^2_{\mathbf{x}}$, one has

$$\partial_\varepsilon E(\mathbf{x}(\cdot) + \varepsilon \boldsymbol{\xi}(\cdot))_{|_{\varepsilon=0}} = \int_0^1 \left(2\left\langle \dot{\mathbf{x}}_t \,,\, \dot{\boldsymbol{\xi}}_t \right\rangle_{\mathbf{x}(t)} + \partial_1 P(\mathbf{x}, \dot{\mathbf{x}}_t)^T \boldsymbol{\xi}\right) dt.$$

Using gradient descent requires selecting an appropriate metric on the space of all time-dependent objects, and interesting developments arise when selecting a metric for which the constraints are continuous functionals [26, 37]. Consider, as an example, the inner product

$$\langle \boldsymbol{\xi}(\cdot)\,,\,\boldsymbol{\eta}(\cdot)\rangle = \int_0^1 \dot{\boldsymbol{\xi}}_t^T \dot{\boldsymbol{\eta}}_t dt$$

that we restrict to the space of time-dependent $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ that vanish at $t = 0$ and $t = 1$. One can check that, defining

$$\begin{aligned}\boldsymbol{\eta}_{\mathbf{x}}(t) = &\int_0^t S_V(\mathbf{x}(u))\dot{\mathbf{x}}_t(u)du - \int_0^t \int_0^u \partial_1 \mathbf{P}(\mathbf{x}(\tilde{u}), \dot{\mathbf{x}}_t(\tilde{u}))d\tilde{u}du \\ &+(1+t)\int_0^t \partial_1 \mathbf{P}(\mathbf{x}(u), \dot{\mathbf{x}}_t(u))du,\end{aligned}$$

one has

$$\partial_\varepsilon E(\mathbf{x}(\cdot) + \varepsilon\boldsymbol{\xi}(\cdot))_{|_{\varepsilon=0}} = \langle \boldsymbol{\xi}\,,\,\nabla E(\mathbf{x})\rangle$$

with

$$\nabla E(\mathbf{x})(t) = \boldsymbol{\eta}_{\mathbf{x}}(t) - t\,\boldsymbol{\eta}_{\mathbf{x}}(1).$$

One can therefore use gradient descent to minimize the geodesic energy, in the form

$$\mathbf{x}^{(n+1)}(t) = \mathbf{x}^{(n)}(t) - \varepsilon(\boldsymbol{\eta}_{\mathbf{x}^{(n)}}(t) - t\,\boldsymbol{\eta}_{\mathbf{x}^{(n)}}(1)).$$

Computing Geodesics Between Curves

We now discuss whether, and how, the previous methods extend to the computation of minimizing geodesics in Riemannian spaces of curves. We start with the metric associated with the projection from 2D diffeomorphisms, since it belongs to the same family as the one discussed with landmarks. In fact, there is a simple way to discretize a curve matching problem so that it boils down to a landmark matching problem. Assume that a reference curve $\mathbf{x}_0$ and a target curve $\mathbf{y}$ are given but that they are only observable in discrete versions, as sequences of points $\mathbf{x}_0^{\text{disc}} = (x_{0,1}, \ldots, x_{0,N})$ and $\mathbf{y}^{\text{disc}} = (y_1, \ldots, y_N)$. Then, as we have remarked, Eq. (25) when restricted to discrete momenta of the form

$$\mu = \sum_{k=1}^N a_k \otimes \delta_{x_k},$$

boils down to Eq. (6), and one can now solve the problem of finding a solution of this equation that transports $x_{0,k}$ to y_k exactly as in the previous sections.

Unfortunately, such an approach has little practical use, given that it is very unlikely that two discrete curves are observed such that the points that constitute them are exactly homologous. This means that one should not require a given $x_{0,k}$ to transform exactly into y_k, but maybe to another y_l or in between two of them. Most of the time, anyway, the curves are given with different numbers of points.

This issue is obviously the discrete form of the parameterization invariance that has been discussed in section "Spaces of Plane Curves." We know that the horizontality condition for parameterization invariance induces the constraint that a_k is perpendicular to the reference curve. In this context, the problem in the continuum is formulated as: given $\mathbf{x}_0$ and $\mathbf{y}$, find an initial momentum μ_0 which is horizontal at $\mathbf{x}_0$ and such that the solution of (25) transforms the curve $\mathbf{x}_0$ into a deformed curve $\varphi(1, \mathbf{x}_0)$ which coincides with $\mathbf{y}$ up to a change of parameterization.

A change of parameter being a diffeomorphism of S^1, it can be generated with an equation like (25). Roughly speaking, this change of parameter can be generated by momenta that are scalar functions on the unit circle. Horizontal geodesics in spaces of curves (still roughly speaking) are generated by momenta that are normal to the reference curve, which can also be represented as scalar functions on the unit disc. So, one needs to find two scalar functions (one for the reparameterization and one for the deformation) that bring the reference curve $\mathbf{x}_0$ to the target $\mathbf{y}$; the target being also characterized by two scalar functions (its coordinates), one sees that the dimensions match and that an approach based on zero finding is possible, at least in principle (there has been no attempt so far in the literature to solve the curve comparison problem in this way). The problem needs to be properly discretized, using, for example, the same number of points to represent $\mathbf{x}_0$, $\mathbf{y}$, the reparameterization momentum, and the deformation momentum.

One can also use a variational approach in the initial momentum, using an objective function like

$$E(\mathbf{a}_0) = d\left(\exp^{\flat}_{\mathbf{x}_0}(\mathbf{a}_0), \mathbf{y}\right)^2 \tag{76}$$

where d is a reparameterization-invariant distance, like the ones in Eqs. (46) and (47), which are, since they derive from Hilbert norms, well amenable to variational computations. The initial momentum a_0 can be discretized as

$$\mathbf{a}_0 = \sum_{k=1}^{N} a_{0,k} \otimes \delta_{u_k},$$

where $u_1, \ldots, u_N$ is a discretization of the unit disc, which, as already noticed, lead to geodesic equations identical to the ones considered with landmarks in (6), the initial "landmark positions" being $x_{0,k} = \mathbf{x}_0(u_k)$. This implies that the variational methods discussed in section "Landmark Matching via Shooting" directly apply, simply changing the objective function.

In fact, the same point of view can also be used with other metrics on curves, with the correct version of the exponential chart or of the momentum representation (whichever is more convenient). Notice that enforcing the fact that the initial momentum is horizontal for reparameterization is optional for these methods, as long as the objective function (the distance d) is parameterization invariant. Disregarding discretization issues, the optimal solution will always be horizontal, so one does not need to make an exact count of the minimal number of degrees of freedom, as was required by zero-finding methods. In practice, the computational efficiency resulting from the reduction of the number of variables can be counterbalanced by the additional flexibility in moving in the space of solutions which is offered by over-parameterized formulations, the choice between the two options being problem dependent.

Finally, notice that path-minimizing methods are also available for curve matching (an approach similar to the one discussed for landmarks in the previous section has been proposed in [37]).

Inexact Matching and Optimal Control Formulation

Inexact Matching

In many cases, requiring an exact representation of the target $\mathbf{y}$ in the exponential chart is not needed and even undesirable. In most instances, indeed, there is an inherent inaccuracy in the way objects are acquired. Landmarks, whether manually or automatically selected, are rarely well defined, and the process can lead to significant variability. The same holds for curves, or surfaces, which are generally extracted using segmentation algorithms, sometimes applied to noisy data, with results that cannot be assumed to be perfect.

Formulations in which geodesics are only required to provide a good approximation of the target then make sense and have a large range of applications. They are akin to the variational methods that were discussed for exact representation, in that they minimize an appropriate distance between the end point of a geodesic and the target, but they also include a penalty term on the length or the energy of the geodesic. In other terms, instead of minimizing $d\left(\exp^{\flat}_{\mathbf{x}_0}(\mathbf{a}_0), \mathbf{y}\right)^2$ like in (76), for example, one would minimize

$$E(\mathbf{a}_0) = d\left(\exp^{\flat}_{\mathbf{x}_0}(\mathbf{a}_0), \mathbf{y}\right)^2 + \sigma^2 \|\mathbf{a}_0\|^2_{\mathbf{x}_0}$$

(in the momentum representation, the norm is for the dual metric in the cotangent space at $\mathbf{x}_0$). If it is more convenient to use an exponential chart instead of the momentum representation, just minimize, over all tangent vectors $\boldsymbol{\xi}_0$ at $\mathbf{x}_0$,

$$E(\boldsymbol{\xi}_0) = d\left(\exp_{\mathbf{x}_0}(\boldsymbol{\xi}_0), \mathbf{y}\right)^2 + \sigma^2 \|\boldsymbol{\xi}_0\|^2_{\mathbf{x}_0}$$

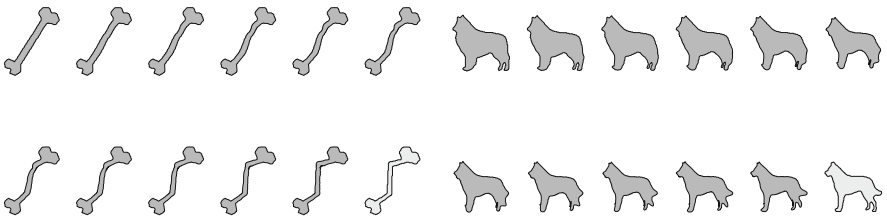


Fig. 6 Two results of inexact matching with the vector-measure distance between curves as error term. The lower right curve (*light gray*) is the target. The first 11 curves provide the geodesic evolution

Since this formulation only adds the term $2\sigma^2\mathbf{a}_0$ (or $2\sigma^2\boldsymbol{\xi}_0$) to the gradient of the objective function that was used for exact representation (i.e., with $\sigma^2 = 0$), the methods that were described in the previous paragraphs can be adapted with minor changes and yield, for example, results like those provided in Fig. 6. Interestingly, in this context, additional methods, deriving from optimal control theory, become available too.

Optimal Control Formulation

Let us first return to the general principles discussed in section "General Principles" and consider an optimal control problem with an additional end-point cost E:

$$\text{minimize} \int_0^1 L(q,u)dt + E(q(1)) \text{ subject to } \dot{q} = f(q,u) \text{ and } q(0) \text{ fixed.}$$

Notice that here $q(1)$ is free, but this situation is handled quite similarly to the one with fixed $q(1)$. Introduce

$$\begin{aligned} J_E(q,p,u) &= J_0(q,p,u) + E(q(1)) \\ &= \int_0^1 \left(L(q,u) + (p|\,\dot{q}_t - f(q,u))\right) dt + E(q(1)). \end{aligned}$$

The only change in the analysis arises when working out the variation in q which now gives the extra end-point condition

$$p(1) + DE(q(1)) = 0\,, \tag{77}$$

which come in addition to the previously obtained (13).

The conservation of the momentum map can be extended to this case when a group G acts on Q and the Hamiltonian H is G-invariant. If, in addition, E is also G-invariant, one deduces from $E(qg) = E(q)$ for all g the fact that $(DE(q)|\,\xi g) = 0$ for all $\xi \in \mathfrak{G}$ which is exactly $\mathfrak{m}(q, DE(q)) = 0$ where the momentum map $\mathfrak{m}$ is

defined in (16). Therefore, (77) implies that $\mathfrak{m}(q(1), p(1)) = 0$, which, combined with the conservation of momentum, implies that

$$\mathfrak{m}(q(t), p(t)) = 0\,.$$

for any $t \in [0, 1]$. Thus, the momentum map is not only invariant but vanishes along solutions of the optimal control problem. In the context of the reduction discussed in section "Reduction: Transitive Group Action," this says that the momentum associated to a solution is horizontal.

Gradient w.r.t. the Control

One can compute the variations with respect to u of

$$C \doteq \int_0^1 L(q, u)dt + E(q(1))$$

subject to the constraint $\dot{q} = f(q, u)$ and $q(0)$ fixed.

Taking the variation with respect to this constraint yields

$$\partial_t \delta q = \partial_q f \delta q + \partial_u f \delta u.$$

Introduce the semigroup $P_{s,t}$ solution of $\partial_t P_{s,t} = \partial_q f P_{s,t}$ with $P_{s,s} = \mathrm{id}$, so that

$$\delta q_t = \int_0^t P_{s,t}(\partial_u f)_s \delta u_s ds.$$

One can write

$$\delta C = \int_0^1 \left(\left((\partial_q L)_t | \int_0^t P_{s,t}(\partial_u f)_s \delta u_s ds \right) + (\partial_u L | \, \delta u(t)) \right) dt$$

$$+ \left(DE(q(1)) | \int_0^1 P_{s,1}(\partial_u f)_s \delta u_s ds \right).$$

Interverting integrals in s and t yields

$$\delta C = \int_0^1 \left(\partial_u L - (\partial_u f)_s^* p(s) | \, \delta u(s)\right) ds = \int_0^1 (-\partial_u H | \, \delta u(s)) ds \tag{78}$$

with

$$p(s) \doteq -\left(\int_s^1 P_{s,t}^* (\partial_q L)_t dt + P_{s,1}^* DE\big(q(1)\big) \right)$$

which is characterized by $p(1) + DE(q(1)) = 0$ and $\partial_t p = \partial_q L - \partial_q f^* p = -\partial_q H(q, p, u)$. The last two conditions are precisely $\delta J_E/\delta q = 0$ for

$$J_E = \int_0^1 ((p|\dot{q}) - H(q, p, u))dt + E(q(1))$$

as above. Since $\dot{q}_t = f(q, u)$ is $\delta J_E/\delta p = 0$, one gets from (78) that $\delta C/\delta u = \delta J_E/\delta u$ for $\delta J_E/\delta p = \delta J_E/\delta q = 0$.

Application to the Landmark Case

In the landmark case $u = \boldsymbol{\alpha}$, $q = \mathbf{x}$, $L(\mathbf{x}, \boldsymbol{\alpha}) = \boldsymbol{\alpha}^T S_V(\mathbf{x})\boldsymbol{\alpha}/2$, and $\dot{\mathbf{x}} = f(\mathbf{x}, \boldsymbol{\alpha}) = S_V(\mathbf{x})\boldsymbol{\alpha}$, so that $\partial_u H(\mathbf{x}, p, \boldsymbol{\alpha}) = S_V(\mathbf{x})(p - \boldsymbol{\alpha})$ and

$$\delta C = \int_0^1 \langle \boldsymbol{\alpha} - p\ ,\ \delta\boldsymbol{\alpha}(s)\rangle_{\mathbf{x}} ds$$

The gradient of C is therefore particularly simple to compute if one chooses along the path the natural metric given on the $\boldsymbol{\alpha}$'s by the matrix $S_V(\mathbf{x})$ (cf. section "A Riemannian Manifold of Deformable Landmarks"). This gives the updating rule (see [22]): $\boldsymbol{\alpha}^{n+1} = \boldsymbol{\alpha}^n - \Delta t(\boldsymbol{\alpha} - p^n)$, $\dot{q}_t^{n+1} = f(q^{n+1}, \boldsymbol{\alpha}^{n+1})$, where p^n is computed by the backward integration of the ode $\dot{p}_t^n = -\partial_q H(q^n, p^n, \boldsymbol{\alpha}^n)$ with end-point condition $p^n(1) + E(q^n(1)) = 0$.

5 Conclusion

Even if it would be impossible to provide a comprehensive description of every method that has been devised in this domain, this chapter provides an introduction to many of the mathematical constructions of spaces of shapes. The combined description of the Riemannian and of the Hamiltonian point of views, which are complementary, should help the reader to a more thorough understanding of the range of available methods, whether they were described in this chapter or elsewhere in the literature. The described numerical methods are basic components that can also be found in most of the contributions that were not directly addressed here.

Mathematical shape analysis remains a domain of intensive research, with open problems arising both for fundamental aspects (e.g., with building spaces of three-dimensional shapes) and for numerical issues and their connections with applications. It is however likely that the concepts introduced here will remain relevant and serve as foundations for future work.

Cross-References

- ▸ Large-Scale Inverse Problems in Imaging
- ▸ Manifold Intrinsic Similarity
- ▸ Variational Methods in Shape Analysis

References

1. Allassonniere, S., Trouve, A., Younes, L.: Geodesic shooting and diffeomorphic matching via textured meshes. In: Proceedings of Energy Minimization Methods in Computer Vision and Pattern Recognition (EMMCVPR), St. Augustine. Volume 3757 of Lecture Notes in Computer Sciences. Springer, Berlin/Heidelberg (2005)
2. Amit, Y., Piccioni, P.: A non-homogeneous Markov process for the estimation of Gaussian random fields with non-linear observations. Ann. Probab. **19**, 1664–1678 (1991)
3. Arad, N., Dyn, N., Reisfeld, D., Yeshurun, Y.: Image warping by radial basis functions: application to facial expressions. CVGIP: Graph. Models Image Process. **56**(2), 161–172 (1994)
4. Arad, N., Reisfeld, D.: Image warping using few anchor points and radial functions. Comput. Graph. Forum **14**, 35–46 (1995)
5. Arnold, V.I.: Sur un principe variationnel pour les ecoulements stationnaires des liquides parfaits et ses applications aux problemes de stanbilite non lineaires. J Mec. **5**, 29–43 (1966)
6. Arnold, V.I.: Mathematical Methods of Classical Mechanics. Springer, New York (1978). Second edition (1989)
7. Aronszajn, N.: Theory of reproducing kernels. Trans. Am. Math. Soc. **68**, 337–404 (1950)
8. Beg, M.F., Miller, M.I., Trouve, A., Younes, L.: Computing large deformation metric mappings via geodesic flows of diffeomorphisms. Int. J. Comput. Vis. **61**(2), 139–157 (2005)
9. Bookstein, F.L.: Principal warps: thin plate splines and the decomposition of deformations. IEEE Trans. Pattern Anal. Mach. Intell. **11**(6), 567–585 (1989)
10. Bookstein, F.L.: Morphometric Tools for Landmark Data: Geometry and Biology. Cambridge University Press, Cambridge (1991)
11. Camion, V., Younes, L.: Geodesic interpolating splines. In: Figueiredo, M., Zerubia, J., Jain, K. (eds.) Proceedings of Energy Minimization Methods in Computer Vision and Pattern Recognition (EMMCVPR), Sophia Antipolis. Volume 2134 of Lecture Notes in Computer Sciences, pp. 513–527. Springer, Berlin (2001)
12. Christensen, G.E., Rabbitt, R.D., Miller, M.I.: Deformable templates using large deformation kinematics. IEEE Trans. Image Process. **5**(10), 1435–1447 (1996)
13. Davis, B.C., Fletcher, P.T., Bullitt, E., Joshi, S.: Population shape regression from random design data. In: IEEE 11th International Conference on Computer Vision (ICCV), Rio de Janeiro, pp. 1–7 (2007)
14. Do Carmo, M.P.: Riemannian Geometry. Birkäuser, Boston (1992)
15. Dryden, I.L., Mardia, K.V.: Statistical Shape Analysis. Wiley, New York (1998)
16. Duchon, J.: Interpolation des fonctions de deux variables suivant le principe de la exion des plaques minces. R.A.I.R.O. Anal. Numer. **10**, 5–12 (1977)
17. Dupuis, P., Grenander, U., Miller, M.: Variational problems on flows of diffeomorphisms for image matching. Q. Appl. Math. **56**, 587–600 (1998)
18. Dyn, N.: Interpolation and approximation by radial and related functions. In: Chui, C.K., Shumaker, L.L., Ward, J.D. (eds.) Approximation Theory VI, vol. 1, pp. 211–234. Academic, San Diego (1989)
19. Federer, H.: Geometric Measure Theory. Springer, New York (1969)
20. Fletcher, P.T., Lu, C., Pizer, M., Joshi, S.: Principal geodesic analysis for the study of nonlinear statistics of shape. IEEE Trans. Med. Imaging **23**(8), 995–1005 (2004)
21. Fletcher, P.T., Venkatasubramanian, S., Joshi, S.: Robust statistics on Riemannian manifolds via the geometric median. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Anchorage, pp. 1–8 (2008)
22. Glaunes, J.: Transport par difféomorphismes de points, de mesures et de courants pour la comparaison de formes et l'anatomie numérique. Ph.D. thesis, University of Paris 13, Paris (in French) (2005)
23. Glaunes, J., Qiu, A., Miller, M.I., Younes, L.: Large deformation diffeomorphic curve matching. Int. J. Comput. Vis. **80**(3), 317–336 (2008)

24. Glaunes, J., Trouve, A., Younes, L.: Diffeomorphic matching of distributions: a new approach for unlabelled point-sets and sub-manifolds matching. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Washington, DC (2004)
25. Glaunes, J., Trouve, A., Younes, L.: Modeling planar shape variation via Hamiltonian flows of curves. In: Krim, H., Yezzi, A. (eds.) Statistics and Analysis of Shapes, pp. 335–361. Springer Birkhauser (2006)
26. Glaunes, J., Vaillant, M., Miller, M.I.: Landmark matching via large deformation diffeomorphisms on the sphere. J. Math. Imaging Vis. **20**, 179–200 (2004)
27. Grenander, U.: General Pattern Theory. Oxford Science Publications, Oxford (1993)
28. Grenander, U., Chow, Y., Keenan, D.M.: Hands: A Pattern Theoretic Study of Biological Shapes. Springer, New York (1991)
29. Grenander, U., Keenan, D.M.: On the shape of plane images. SIAM J. Appl. Math. **53**(4), 1072–1094 (1991)
30. Grenander, U., Miller, M.I.: Computational anatomy: an emerging discipline. Q. Appl. Math. **LVI**(4), 617–694 (1998)
31. Hamilton, R.S.: The inverse function theorem of Nash and Moser. Bull. Am. Math. Soc. (N.S.) **7**(1), 65–222 (1982)
32. Helgason, S.: Differential Geometry, Lie Groups and Symmetric Spaces. Academic, New York (1978)
33. Holm, D.D.: Geometric Mechanics. Imperial College Press, London (2008)
34. Holm, D.D., Marsden, J.E., Ratiu, T.S.: The Euler–Poincaré equations and semidirect products with applications to continuum theories. Adv. Math. **137**, 1–81 (1998)
35. Holm, D.R., Trouvé, A., Younes, L.: The Euler–Poincaré theory of metamorphosis. Q. Appl. Math. **67**, 661–685 (2009)
36. Joshi, S., Miller, M.: Landmark matching via large deformation diffeomorphisms. IEEE Trans. Image Process. **9**(8), 1357–1370 (2000)
37. Joshi, S.H., Klassen, E., Srivastava, A., Jermyn, I.: A novel representation for Riemannian analysis of elastic curves in Rn. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Minneapolis (2007)
38. Jost, J.: Riemannian Geometry and Geometric Analysis, 2nd edn. Springer, Berlin (1998)
39. Karcher, H.: Riemannian center of mass and mollifier smoothing. Commun. Pure Appl. Math. **30**(5), 509–541 (1977)
40. Kendall, D.G.: Shape manifolds, Procrustean metrics and complex projective spaces. Bull. Lond. Math. Soc. **16**, 81–121 (1984)
41. Kendall, D.G., Barden, D., Carne, T.K., Le, H.: Shape and Shape Theory. Wiley, New York (1999)
42. Klassen, E., Srivastava, A., Mio, W., Joshi, S.: Analysis of planar shapes using geodesic paths on shape spaces. IEEE Trans. Pattern Anal. Mach. Intell. **24**, 375–405 (2002)
43. Klassen, E., Srivastava, A., Mio, W., Joshi, S.H.: Analysis of planar shapes using geodesic paths on shape spaces. IEEE Trans. Pattern Anal. Mach. Intell. **26**(3), 372–383 (2004)
44. Kriegl, A., Michor, P.W.: The Convenient Setting of Global Analysis. Mathematical Surveys and Monographs, vol. 53. AMS, Providence (1997)
45. Kriegl, A., Michor, P.W.: Regular infinite dimensional lie groups. J. Lie Theory **7**(1), 61–99 (1997)
46. Le, H.: Mean size-and-shapes and mean shapes: a geometric point of view. Adv. Appl. Probl. **27**, 44–55 (1995)
47. Le, H.: Estimation of Riemannian barycentres. Lond. Math. Soc. J. Comput. Math. **7**, 193–200 (2004)
48. Marques, J.A., Abrantes, A.J.: Shape alignment-optimal initial point and pose estimation. Pattern Recognit. Lett. **18**, 49–53 (1997)
49. Marsden, J.E.: Lectures on Geometric Mechanics. Cambridge University Press, New York (1992)
50. Marsden, J.E., Ratiu, T.S.: Introduction to Mechanics and Symmetry. Springer, Berlin (1999)

51. Meinguet, J.: Multivariate interpolation at arbitrary points made simple. J. Appl. Math. Phys. **30**, 292–304 (1979)
52. Mennucci, A., Yezzi, A.: Metrics in the space of curves. Technical report, arXiv:mathDG/0412454 v2 (2005)
53. Micheli, M.: The differential geometry of landmark shape manifolds: metrics, geodesics, and curvature. Ph.D. thesis, Brown University, Providence (2008)
54. Michor, P.W., Mumford, D.: Vanishing geodesic distance on spaces of submanifolds and diffeomorphisms. Doc. Math. **10**, 217–245 (2005)
55. Michor, P.W., Mumford, D.: Riemannian geometries on spaces of plane curves. J. Eur. Math. Soc. **8**, 1–48 (2006)
56. Michor, P.W., Mumford, D.: An overview of the Riemannian metrics on spaces of curves using the Hamiltonian approach. Appl. Comput. Harmonic Anal. **23**(1), 74–113 (2007)
57. Miller, M.I., Trouvé, A., Younes, L.: Geodesic shooting for computational anatomy. J. Math. Image Vis. **24**(2), 209–228 (2006)
58. Miller, M.I., Younes, L.: Group action, diffeomorphism and matching: a general framework. Int. J. Comput. Vis. **41**, 61–84 (2001). (Originally published in electronic form in: Proceeding of SCTV 99, http://www.cis.ohiostate.edu/szhu/SCTV99.html)
59. O'Neill, B.: The fundamental equations of a submersion. Mich. Math. J. **13**, 459–469 (1966)
60. Qiu, A., Younes, L., Miller, M.I.: Intrinsic and extrinsic analysis in computational anatomy. NeuroImage **39**(4), 1804–1814 (2008)
61. Qiu, A., Younes, L., Wang, L., Ratnanather, J.T., Gillepsie, S.K., Kaplan, K., Csernansky, J., Miller, M.I.: Combining anatomical manifold information via diffeomorphic metric mappings for studying cortical thinning of the cingulate gyrus in schizophrenia. NeuroImage **37**(3), 821–833 (2007)
62. Shah, J.: H^0 type Riemannian metrics on the space of planar curves. Q. Appl. Math. **66**, 123–137 (2008)
63. Sharon, E., Mumford, D.: 2D-shape analysis using conformal mapping. Int. J. Comput. Vis. **70**(1), 55–75 (2006)
64. Small, C.: The statistical Theory of Shape. Springer, New York (1996)
65. Thompson, D.W.: On Growth and Form. Dover, Mineola (1917). Revised edition (1992)
66. Trouvé, A.: Action de groupe de dimension infinie et reconnaissance de formes. C. R. Acad. Sci. Paris Ser. I Math. **321**(8), 1031–1034 (1995)
67. Trouvé, A.: Diffeomorphism groups and pattern matching in image analysis. Int. J. Comput. Vis. **28**(3), 213–221 (1998)
68. Trouvé, A., Younes, L.: Diffeomorphic matching in 1D: designing and minimizing matching functionals. In: Vernon, D. (ed.) Proceedings of European Conference on Computer Vision (ECCV), Dublin (2000)
69. Trouvé, A., Younes, L.: On a class of optimal matching problems in 1 dimension. SIAM J. Control Opt. **39**(4), 1112–1135 (2001)
70. Trouvé, A., Younes, L.: Local geometry of deformable templates. SIAM J. Math. Anal. **37**(1), 17–59 (2005)
71. Trouvé, A., Younes, L.: Metamorphoses through lie group action. Found. Comput. Math. **5**, 173–198 (2005)
72. Twinings, C., Marsland, S., Taylor, C.: Measuring geodesic distances on the space of bounded diffeomorphisms. In: British Machine Vision Conference, Cardiff (2002)
73. Vaillant, M., Glaunés, J.: Surface matching via currents. In: Christensen, G.E., Milan S. (eds.) Proceedings of Information Processing in Medical Imaging (IPMI), Glenwood Springs. Volume 3565 in Lecture Notes in Computer Science. Springer (2005)
74. Vaillant, M., Miller, M.I., Trouvé, A., Younes, L.: Statistics on diffeomorphisms via tangent space representations. NeuroImage **23**(S1), S161–S169 (2004)
75. Vialard F.-X.: Hamiltonian approach to shape spaces in a diffeomorphic framework: from the discontinuous image matching problem to a stochastic growth model. Ph.D. thesis, Ecole Normale Supérieure de Cachan. http://tel.archives-ouvertes.fr/tel-00400379/fr/ (2009)

76. Vialard F.-X., Santambrogio, F.: Extension to BV functions of the large deformation diffeomorphisms matching approach. C. R. Math. **347**(1–2), 27–32 (2009)
77. Wahba, G.: Spline Models for Observational Data. SIAM, Philadelphia (2006)
78. Wang, L., Beg, M.F., Ratnanather, J.T., Ceritoglu, C., Younes, L., Morris, J., Csernansky, J., Miller, M.I.: Large deformation diffeomorphism and momentum based hippocampal shape discrimination in dementia of the Alzheimer type. IEEE Trans. Med. Imaging **26**, 462–470 (2006)
79. Younes, L.: Computable elastic distances between shapes. SIAM J. Appl. Math. **58**(2), 565–586 (1998)
80. Younes, L.: Optimal matching between shapes via elastic deformations. Image Vis. Comput. **17**, 381–389 (1999)
81. Younes, L., Michor, P., Shah, J., Mumford, D.: A metric on shape spaces with explicit geodesics. Rend. Lincei Math. Appl. **9**, 25–57 (2008)

Variational Methods in Shape Analysis

Martin Rumpf and Benedikt Wirth

Contents

1 Introduction 1820
2 Background 1820
3 Mathematical Modeling and Analysis 1824
Recalling the Finite-Dimensional Case 1824
Path-Based Viscous Dissipation Versus State-Based Elastic Deformation for Non-rigid Objects 1827
4 Numerical Methods and Case Examples 1835
Elasticity-Based Shape Space 1835
Viscous Fluid-Based Shape Space 1843
A Collection of Computational Tools 1850
5 Conclusion 1853
Cross-References 1854
References 1855

Abstract

The concept of a shape space is linked both to concepts from geometry and from physics. On one hand, a path-based viscous flow approach leads to Riemannian distances between shapes, where shapes are boundaries of objects that mainly behave like fluids. On the other hand, a state-based elasticity approach induces a (by construction) non-Riemannian dissimilarity measure between shapes, which is given by the stored elastic energy of deformations matching the corresponding objects. The two approaches are both based on variational principles. They are analyzed with regard to different applications, and a detailed comparison is given.

M. Rumpf (✉) • B. Wirth
Institute for Numerical Simulation, Bonn University, Bonn, Germany
e-mail: martin.rumpf@ins.uni-bonn.de; benedikt.wirth@ins.uni-bonn.de

© Springer Science+Business Media New York 2015

O. Scherzer (ed.), *Handbook of Mathematical Methods in Imaging*,
DOI 10.1007/978-1-4939-0790-8_56

1 Introduction

The analysis of shapes as elements in a frequently infinite-dimensional space of shapes has attracted increasing attention over the last decade. There are pioneering contributions in the theoretical foundation of shape space as a Riemannian manifold as well as path-breaking applications to quantitative shape comparison, shape recognition, and shape statistics. The aim of this chapter is to adopt a primarily physical perspective on the space of shapes and to relate this to the prevailing geometric perspective. Indeed, we here consider shapes given as boundary contours of volumetric objects, which consist either of a viscous fluid or an elastic solid.

In the first case, shapes are transformed into each other via viscous transport of fluid material, and the flow naturally generates a connecting *path* in the space of shapes. The viscous dissipation rate – the rate at which energy is converted into heat due to friction – can be defined as a metric on an associated Riemannian manifold. Hence, via the computation of the shortest transport paths, one defines a distance measure between shapes.

In the second case, shapes are transformed via elastic deformations, where the associated elastic energy only depends on the final *state* of the deformation and not on the path along which the deformation is generated. The minimal elastic energy required to deform an object into another one can be considered as a dissimilarity measure between the corresponding shapes.

In what follows, we discuss and extensively compare the *path*-based and the *state*-based approach. As applications of the elastic shape model, we consider shape averages and a principal component analysis of shapes. The viscous flow model is used to exemplarily cluster 2D and 3D shapes and to construct a flow-type nonlinear interpolation scheme. Furthermore, we show how to approximate the viscous, path-based approach with a time-discrete sequence of state-based variational problems.

2 Background

The structure of shape spaces and statistical analyses of shapes have been examined in various settings, and applications range from the computation of priors for segmentation [16, 17, 43] and shape classification [25, 44, 48, 50] to the construction of standardized anatomical atlases [14, 37, 66]. Among all existing approaches, a number of different concepts of a shape are employed, including landmark vectors [16, 39], planar curves [41, 52, 84], surfaces in $\mathbb{R}^3$ [24, 25, 40], boundary contours of objects [31, 44, 67], multiphase objects [83], as well as the morphologies of images [22].

The analysis of a shape space is typically based on a notion of a distance or dissimilarity measure $d(\cdot, \cdot)$ between shapes [10, 31, 50, 51, 54, 67], whose definition frequently takes a variational form. This distance can be used to define an average [26, 67] or a median [4, 28] $\mathcal{S}$ of given shapes $\mathcal{S}_1, \ldots, \mathcal{S}_n$ according to $\mathcal{S} = \operatorname{argmin}_{\tilde{\mathcal{S}}} \sum_{i=1}^{n} d(\tilde{\mathcal{S}}, \mathcal{S}_i)^p$ for $p = 1$ and $p = 2$, respectively (cf.

section "Elastic Shape Averaging"). Likewise, shape variations can be obtained by a principal component analysis (PCA, cf. section "Elasticity-Based PCA") or a more general covariance analysis in a way which is consistent with the dissimilarity measure between shapes [11, 16, 26, 68]. From the conceptional point of view, one can distinguish two types of these dissimilarities or distance measures which may be characterized as rather state based or path based, respectively. While the first approach is independent of the notion of paths of shapes, the latter distance definition requires the computation of an optimal, connecting path in shape space. In some cases, both concepts coincide: The Euclidean distance between two points, e.g., can equivalently be interpreted in a state-based manner as the norm of the difference vector or as the length of the shortest connecting path (we shall provide a physical interpretation for each case in section "Recalling the Finite-Dimensional Case").

The notion of a shape space was already introduced by Kendall in 1984 [39], who considers shapes as k-tuples of points in $\mathbb{R}^d$, endowed with the quotient metric of $\mathbb{R}^{kd}$ with respect to similarity transforms. Often, however, a shape space is just modeled as a linear vector space which is not invariant with respect to shift or rotation a priori. In the simplest case, such a shape space is made up of vectors of landmark positions, and distances between shapes can be evaluated in a state-based manner as the Euclidean norm of their difference. Chen and Parent [12] investigated averages of 2D contours already in 1989. Cootes et al. perform a PCA on training shapes with consistently placed landmarks to obtain priors for edge-based image segmentation [16]. Hafner et al. use a PCA of position vectors covering the proximal tibia to reconstruct the tibia surface just from six dominant modes [35]. Perperidis et al. automatically assign consistent landmarks to training shapes by a nonrigid registration as a preprocessing step for a PCA of the cardiac anatomy [63]. Söhn et al. compute dominant eigenmodes of landmark displacement on human organs, also using registration for preprocessing [73].

As an infinite-dimensional vector space, the Lebesgue-space L^2 has served as shape space, where again shape alignment is a necessary preprocessing step. Leventon et al. identify shapes with their signed distance functions and impose the Hilbert space structure of L^2 on them to compute an average and dominant modes of variation [43]. Tsai et al. apply the same technique to 3D prostate images [79]. Dambreville et al. also compute shape priors, but using characteristic instead of signed distance functions [19].

A more sophisticated state-based shape space is obtained by considering shapes as subsets of an ambient space with a metric $d(\cdot, \cdot)$ and endowing them with the Hausdorff distance

$$d_{\mathrm{H}}(\mathcal{S}_1, \mathcal{S}_2) = \max\{\sup_{x \in \mathcal{S}_1} \inf_{y \in \mathcal{S}_2} d(x, y), \sup_{y \in \mathcal{S}_1} \inf_{x \in \mathcal{S}_2} d(x, y)\}$$

between any two shapes $\mathcal{S}_1, \mathcal{S}_2$. Charpiat et al. employ smooth approximations of the Hausdorff distance based on a comparison of the signed distance functions of shapes [10]. For a given set of shapes, the gradient of the shape distance functional at the average shape is regarded as shape variation of the average and used to

analyze its dominant modes of variation [11]. Frame indifference is mimicked by an inner product that weights rotations, shifts, scalings, and the orthogonal complement to these transformations differently. Charpiat et al. also consider gradient flow morphing from one shape onto another one which can be regarded as a means to obtain meaningful paths even in shape spaces with state-based distance measures.

An isometrically invariant distance measure between shapes (or more general metric spaces) that is also not based on connecting paths is provided by the Gromov–Hausdorff distance, which can be defined variationally as

$$d_{\mathrm{GH}}(\mathcal{S}_1, \mathcal{S}_2) = \frac{1}{2} \inf_{\substack{\phi:\mathcal{S}_1\to\mathcal{S}_2 \\ \psi:\mathcal{S}_2\to\mathcal{S}_1}} \sup_{\substack{y_i=\phi(x_i) \\ \psi(y_i)=x_i}} |d_{\mathcal{S}_1}(x_1, x_2) - d_{\mathcal{S}_2}(y_1, y_2)|\,,$$

where $d_{\mathcal{S}_i}(\cdot\,,\cdot)$ is a distance measure between points in $\mathcal{S}_i$. The Gromov–Hausdorff distance represents a global, supremum-type measure of the lack of isometry between two shapes. Memoli and Sapiro use this distance for clustering shapes described by point clouds, and they discuss efficient numerical algorithms to compute Gromov–Hausdorff distances based on a robust notion of intrinsic distances $d_{\mathcal{S}}(\cdot\,,\cdot)$ on the shapes [50]. Bronstein et al. incorporate the Gromov–Hausdorff distance concept in various classification and modeling approaches in geometry processing [6]. Memoli investigates the relation between the Gromov–Hausdorff distance and the Hausdorff distance under the action of Euclidean isometries as well as L^p-type variants of the Gromov–Hausdorff distance [49].

In [46], Manay et al. define shape distances via integral invariants of shapes and demonstrate the robustness of this approach with respect to noise.

Another distance or dissimilarity measure which also measures the lack of isometry between shapes can be obtained by interpreting shapes as boundaries of physical objects and measuring the (possibly nonlinear) deformation energy of an elastic matching deformation ϕ between two objects [36, 67]. Since, by the axiom of elasticity, this energy solely depends on the original and the final configuration of the deformed object but not on the deformation path, the elastic dissimilarity measure can clearly be classified as state based (as will be detailed in section "State-Based, Path-Independent Elastic Setup"). This physical approach comes along with a natural linearization of shapes via boundary stresses to perform a covariance analysis [68] and will be presented in section "Elasticity-Based Shape Space." Pennec et al. define a nonlinear elastic energy as the integral over the ambient space of an energy density that depends on the logarithm of the Cauchy–Green strain tensor $\mathcal{D}\phi^T\mathcal{D}\phi$ [61, 62], which induces a symmetric state-based distance.

Typical path-based shape spaces have the structure of a Riemannian manifold. Here, the strength of a shape variation is measured by a Riemannian metric, and the square root of the Riemannian metric evaluated on the temporal shape variation is integrated along a path of shapes to yield the path length. The length of the shortest path between two shapes represents their geodesic distance $d(\cdot\,,\cdot)$. Averages are obtained via the Fréchet mean [30], which was further analyzed by Karcher [38].

There is also a natural linear representation of shapes in the tangent space at the Fréchet mean via the logarithmic map, which enables a PCA.

A Riemannian shape space which might still be regarded as rather state than path oriented is given by the space of polygonal medial axis representations, where each shape is described by a polygonal lattice and spheres around each vertex [87]: Here, the Lie group structure of the medial representation space can be exploited to approximate the Fréchet mean as exponential map of the average of the logarithmic maps of the input. Fletcher et al. perform a PCA on these log-maps to obtain the dominant geometric variations of kidney shapes [26] and brain ventricles [27]. Fuchs and Scherzer use the PCA on log-maps to obtain the covariance of medial representations, and they use a covariance-based Mahalanobis distance to impose a new metric on the shape manifold. This metric is employed to obtain priors for edge-based image segmentation [32, 33].

Kilian et al. compute and extrapolate geodesics between triangulated surfaces of fixed mesh topology, using isometry invariant Riemannian metrics that measure the local distortion of the grid [40]. Eckstein et al. employ different metrics in combination with a smooth approximation to the Hausdorff distance to perform gradient flows for shape matching [24]. Liu et al. use a discrete exterior calculus approach on simplicial complexes to compute geodesics and geodesic distances in the space of triangulated shapes, in particular taking care of higher genus surfaces [45].

An infinite-dimensional Riemannian shape space has been developed for planar curves. Klassen et al. propose to use as a Riemannian metric the L^2-metric on variations of the direction or curvature functions of arc length-parameterized curves. They implement a shooting method to find geodesics [41], while Schmidt and Cremers present an alternative variational approach [70]. Srivastava et al. assign different weights to the L^2-metric on stretching and on bending variations and obtain an elastic model of curves [75]. Michor and Mumford examine Riemannian metrics on the manifold of smooth regular curves [51]. They show the standard L^2-metric in tangent space, leading to arbitrarily short geodesics and hence employ a curvature-weighted L^2-metric instead. Yezzi and Mennucci resolved the problem taking into account the conformal factor in the metric [84]. Sundaramoorthi et al. use Sobolev metrics in the tangent space of planar curves to perform gradient flows for image segmentation via active contours [76]. Michor et al. discuss a specific metric on planar curves, for which geodesics can be described explicitly [52]. In particular, they demonstrate that the sectional curvature on the underlying shape space is bounded from below by zero, which points out a close relation to conjugate points in shape space and thus to only locally the shortest geodesics. Finally, Younes considers a left-invariant Riemannian distance between planar curves by identifying shapes with elements of a Lie group acting on one reference shape [85].

When warping objects bounded by shapes in $\mathbb{R}^d$, a shape tube in $\mathbb{R}^{d+1}$ is formed. Delfour and Zolésio [20] rigorously develop the notion of a Courant metric in this context. A further generalization to classes of non-smooth shapes and the derivation of the Euler–Lagrange equations for a geodesic in terms of a shortest shape tube is investigated by Zolésio in [88].

Dupuis et al. [23] and Miller et al. [53, 54] define the distance between shapes based on a flow formulation in the embedding space. They exploit the fact that in case of sufficient Sobolev regularity for the motion field v on the whole surrounding domain Ω, the induced flow consists of a family of diffeomorphisms. This regularity is ensured by a functional $\int_0^1 \int_\Omega Lv \cdot v \, dx \, dt$, where L is a higher-order elliptic operator [76, 85]. Geometrically, $\int_\Omega Lv \cdot v \, dx$ is the underlying Riemannian metric, and we will discuss related, path-based concepts in section "Path-Based, Viscous Riemannian Setup." Under sufficient smoothness assumptions, Beg et al. derive the Euler–Lagrange equations for the diffeomorphic flow field [3]. To compute geodesics between hypersurfaces in the flow of diffeomorphism framework, a penalty functional measures the distance between the transported initial shape and the given end shape. Vaillant and Glaunès [80] identify hypersurfaces with naturally associated two forms and used the Hilbert space structures on the space of these forms to define a mismatch functional. The case of planar curves is investigated under the same perspective by Glaunès et al. in [34]. To enable the statistical analysis of shape structures, parallel transport along geodesics is proposed by Younes et al. [86] as the suitable tool to transfer structural information from subject-dependent shape representations to a single-template shape.

In most applications, shapes represent boundary contours of physical objects. Fletcher and Whitaker adopt this viewpoint to develop a model for geodesics in shape space which avoids overfolding [29]. Fuchs et al. [31] propose a Riemannian metric on a space of shape contours, motivated by linearized elasticity. This metric can be interpreted as the rate of physical dissipation during the deformation of a viscous liquid object [82, 83] and will be elaborated in section "Viscous Fluid-Based Shape Space."

Finally, a shape space is sometimes understood as a manifold, learned from training shapes, and embedded in a higher-dimensional (often linear) space. Many related approaches are based on kernel density estimation in feature space. Here, the manifold is described by a probability distribution in the embedding space, which is computed by mapping points of the embedding space into a higher-dimensional feature space and assuming a Gaussian distribution there. In general, points in feature space have no exact preimage in shape space, so that approximate preimages have to be obtained via a variational formulation [64]. Cremers et al. use this technique to obtain 2D silhouettes of 3D objects as priors for image segmentation [17]. Rathi et al. provide a comparison between kernel PCA, local linear embedding (LLE), and kernel LLE (kernel PCA only on the nearest neighbors) [65]. Thorstensen et al. approximate the shape manifold using weighted Karcher means of the nearest neighbor shapes obtained by diffusion maps [77].

3 Mathematical Modeling and Analysis

Recalling the Finite-Dimensional Case

At first, let us investigate distances and their relation to concepts from physics in the simple case of Euclidian space. In Euclidean space, the shortest paths are

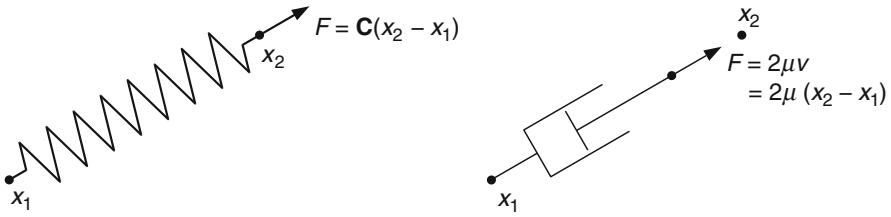


Fig. 1 The force F of an elastic spring between x_1 and x_2 is proportional to $(x_2 - x_1)$, as well as the force F of a dashpot which is extended from x_1 to x_2 within time 1 at constant velocity v. The spring energy reads $\mathcal{W} = \int F \, \mathrm{d}x = \frac{1}{2}\mathbf{C} \, \|x_2 - x_1\|_2^2$ and the dashpot dissipation $\mathbf{Diss} = \int F \cdot v \, \mathrm{d}t = 2\mu \, \|x_2 - x_1\|_2^2$

straight lines, and they are unique, so that the distance computation involves only the states of the two end points: The geodesic distance between any two points $x_1, x_2 \in \mathbb{R}^d$ is given by the norm of the difference, $\|x_2 - x_1\|_2$, which implies the equivalence of the state-based and the path-based perspective. A corresponding physical view might be the following. Considering that – by Hooke's law – the stored elastic energy of an elastic spring extended from x_1 to x_2 is given by $\mathcal{W} = \frac{1}{2}\mathbf{C} \, \|x_2 - x_1\|_2^2$ for the spring constant $\mathbf{C}$, the distance can be interpreted in a state-based manner as the square root of the elastic spring energy (Fig. 1). Likewise, from a path-based point of view, the minimum dissipated energy of a dashpot which is extended from x_1 to x_2 at constant speed within the fixed time interval $[0, 1]$ reads $\mathbf{Diss} = \int_0^1 2\mu \, \|v\|_2^2 \, \mathrm{d}t = 2\mu \, \|x_2 - x_1\|_2^2$, where 2μ is the dashpot parameter and the velocity is given by $v = x_2 - x_1$. Using this physical interpretation, we can express, for instance, the arithmetic mean $x = \frac{1}{n}\sum_{i=1}^n x_i = \operatorname{argmin}_{\tilde{x}} \sum_{i=1}^n \|x_i - \tilde{x}\|_2^2$ of a given set of points $x_1, \ldots, x_n \in \mathbb{R}^d$ either as the minimizer of the total elastic deformation energy in a system, where the average x is connected to each x_i by elastic springs or as the minimizer of the total viscous dissipation when extending dashpots from x_i to x.

Before we investigate the same concepts on more general Riemannian manifolds, let us briefly recall some basic notation. A Riemannian manifold is a set $\mathcal{M}$ that is locally diffeomorphic to Euclidean space. Given a smooth path $x(t) \in \mathcal{M}$, $t \in [0, 1]$, we can define its derivative $\dot{x}(t)$ at time t as a tangent vector to $\mathcal{M}$ at $x(t)$. The vector space of all such tangent vectors makes up the tangent space $T_{x(t)}\mathcal{M}$, and it is equipped with the metric $g_{x(t)}(\cdot, \cdot)$ as the inner product. The length of a path $x(t) \in \mathcal{M}$, $t \in [0, 1]$, is defined as $\int_0^1 \sqrt{g_{x(t)}(\dot{x}(t), \dot{x}(t))} \, \mathrm{d}t$, and locally the shortest paths are denoted geodesics. They can be shown to minimize $\int_0^1 g_{x(t)}(\dot{x}(t), \dot{x}(t)) \, \mathrm{d}t$ [21, Lemma 2.3]. Let us emphasize that a general geodesic is only locally the shortest curve. In particular, there might be multiple geodesics of different lengths connecting the same end points. The geodesic distance between two points is the length of the shortest connecting path. Finally, for a given $x \in \mathcal{M}$, there is a bijection $\exp_x : T_x\mathcal{M} \to \mathcal{M}$ of a neighborhood of $0 \in T_x\mathcal{M}$ into a neighborhood of $x \in \mathcal{M}$ that assigns to each tangent vector $v \in T_x\mathcal{M}$ the end point

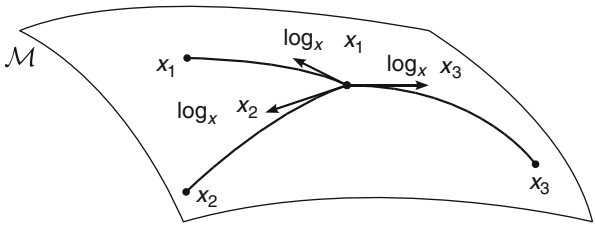


Fig. 2 The logarithmic map assigns each point x_i on the manifold $\mathcal{M}$ a vector in the tangent space $T_x\mathcal{M}$, which may be seen as a linear representative

of the geodesic emanating from x with initial velocity v and running over the time interval $[0, 1]$ [42, Theorem 1.6.12] or [74, Chap. 9, Theorem 14].

We can now define the (possibly nonunique, cf. Sect. 5) mean x of a number of n points $x_1, \ldots, x_n \in \mathcal{M}$ in analogy to the Euclidian case as $x = \operatorname{argmin}_{\tilde{x}} \sum_{i=1}^{n} d(x_i, \tilde{x})^2$, where $d(\cdot, \cdot)$ is the Riemannian distance on $\mathcal{M}$. This average is uniquely defined as long as the geodesics involved in the distance computation are unique, and it has been investigated in differential geometry by Karcher [38]. Furthermore, on a Riemannian manifold $\mathcal{M}$, the inverse exponential map $\log_x = \exp_x^{-1}$ provides a method to obtain representatives $\log_x(x_i) \in T_x\mathcal{M}$ of given input points $x_i \in \mathcal{M}$ in the (linear) vector space $T_x\mathcal{M}$ (Fig. 2). On these, we can perform a PCA, which is by definition a linear statistical tool.

In a Riemannian space $\mathcal{M}$, the path-based approach can immediately be applied by exploiting the Riemannian structure, and $\int_0^1 g_{x(t)}(\dot{x}(t), \dot{x}(t))\,\mathrm{d}t$ can be considered as the energy dissipation spent to move a point from $x(0)$ to $x(1)$ along a geodesic. The logarithms $\log_x(x_i)$ in this model correspond to the initial velocities of the transport process leading from x to x_i. When applying the state-based elastic model in $\mathcal{M}$, however, there is no mechanically motivated notion of paths and thus also no logarithmic map. Only if we suppose that the Riemannian structure of the space $\mathcal{M}$ is not induced by changes in the inner structure of our objects, the physical model based on elastic springs still coincides with the viscous model: We consider elastic springs stretched on the surface $\mathcal{M}$ and connecting the points x and x_i with a stored energy $\frac{1}{2}\mathbf{C}d(x, x_i)^2$. Then, as before in the Euclidian case, a state-based average x of input points $x_1, \ldots, x_n$ can be defined. Furthermore, interpreting spring forces acting on x and pointing toward x_i as linear representatives of the input points x_i, one can run a PCA on these forces as well. However, for any reasonable (even finite-dimensional) model of shape space, objects are not rigid, and the inner relation between points as subunits (such as the vertex points of polygonal shapes) essentially defines the Riemannian (and thus the path-based) structure of the space $\mathcal{M}$: The rate of dissipation along a path in shape space depends on the interaction of object points. Physically, the corresponding point interaction energy is converted into thermal energy via friction. This dissipation depends significantly on the path in shape space traversed from one shape to the other. In contrast, when applying the state-based approach to the same shape space, we directly compare the inner relations between the subunits, i.e., we have no history of these relations. This comparison can be quantified based on a stored (elastic) interaction energy which is

then a quantitative measure of the dissimilarity of the two objects but in general no metric distance.

Path-Based Viscous Dissipation Versus State-Based Elastic Deformation for Nonrigid Objects

In the following, we will especially consider two different physically motivated perspectives on a shape space of nonrigid volumetric objects in more detail. In the first case, we will adopt a path-based view, motivated by the theory of viscous fluids, while the second, state-based approach will be motivated by elasticity.

We will regard shapes $\mathcal{S}$ as boundaries $\mathcal{S} = \partial\mathcal{O}$ of domains $\mathcal{O} \subset \mathbb{R}^d$ which will be interpreted as physical objects. The resulting shape space structure depends on the particular type of physical objects $\mathcal{O}$: An interpretation of $\mathcal{O}$ as a blob of a viscous fluid will yield an actually Riemannian, path-based shape space, while the interpretation as an elastic solid results in a state-based perspective, which will turn out to be non-Riemannian by construction.

Path-Based, Viscous Riemannian Setup

Shapes will be modeled as the boundary contour of a physical object that is made of a viscous fluid. The object might be surrounded by a different fluid (e.g., with much lower viscosity and compression modulus), nevertheless, without any restriction we will assume void outside the object in the derivation of our model. Here, *viscosity* describes the internal resistance in a fluid and is a macroscopic measure of the friction between fluid particles, e.g., the viscosity of honey is significantly larger than that of water. The friction is described in terms of the stress tensor $\sigma = (\sigma_{ij})_{ij=1,\dots d}$, whose entries describe a force per area element. By definition, σ_{ij} is the force component along the ith coordinate direction acting on the area element with a normal pointing in the jth coordinate direction. Hence, the diagonal entries of the stress tensor σ refer to normal stresses, e.g., due to compression, and the off-diagonal entries represent tangential (shear) stresses. The Cauchy stress law states that due to the preservation of angular momentum, the stress tensor σ is symmetric [13].

In a *Newtonian fluid*, the stress tensor is assumed to depend linearly on the gradient $\mathcal{D}v := \left(\frac{\partial v_i}{\partial x_j}\right)_{ij=1,\dots d}$ of the velocity v. In case of a rigid body motion, the stress vanishes. A rotational component of the local motion is generated by the antisymmetric part $\frac{1}{2}\left(\mathcal{D}v - (\mathcal{D}v)^T\right)$ of the velocity gradient, and it has the local rotation axis $\nabla \times v$ and local angular velocity $|\nabla \times v|$ [78]. Thus, as rotations are rigid body motions, the stress only depends on the symmetric part $\epsilon[v] := \frac{1}{2}\left(\mathcal{D}v + (\mathcal{D}v)^T\right)$ of the velocity gradient. For an isotropic Newtonian fluid, we get $\sigma_{ij} = \lambda\delta_{ij}\sum_k(\epsilon[v])_{kk} + 2\mu\,(\epsilon[v])_{ij}$, or in matrix notation $\sigma = \lambda \mathrm{tr}\,(\epsilon[v])\,\mathbb{1} + 2\mu\epsilon[v]$, where $\mathbb{1}$ is the identity matrix. The parameter λ is denoted Lamé's first coefficient. The local rate of viscous dissipation – the rate at which mechanical energy is locally converted into heat due to friction – can now be computed as

$$\mathbf{diss}[v] = \frac{\lambda}{2}(\mathrm{tr}\epsilon[v])^2 + \mu \mathrm{tr}\left(\epsilon[v]^2\right) . \tag{1}$$

This is in direct correspondence to the mechanical definition of the stress tensor σ as the first variation of the local dissipation rate with respect to the velocity gradient, i.e., $\sigma = \delta_{Dv}\mathbf{diss}$. Indeed, by a straightforward computation, we obtain $\delta_{(Dv)_{ij}}\mathbf{diss} = \lambda\, \mathrm{tr}\epsilon[v]\, \delta_{ij} + 2\mu\ (\epsilon[v])_{ij} = \sigma_{ij}$. Here, $\mathrm{tr}\left(\epsilon[v]^2\right)$ measures the averaged local change of length and $(\mathrm{tr}\epsilon[v])^2$ the local change of volume induced by the transport. Obviously, $\mathrm{div}v = \mathrm{tr}(\epsilon[v]) = 0$ characterizes an incompressible fluid.

Now, let us consider a path $(\mathcal{O}(t))_{t\in[0,1]}$ of objects connecting $\mathcal{O}(0)$ with $\mathcal{O}(1)$ and generated by a time-continuous deformation. If each point $x \in \mathcal{O}(t)$ of the object $\mathcal{O}(t)$ at time $t \in [0, 1]$ moves in an Eulerian framework at the velocity $v(t, x)$ ($\dot{x} = v(t, x)$), so that the total deformation of $\mathcal{O}(0)$ into $\mathcal{O}(t)$ can be obtained by integrating the velocity field v in time, then the accumulated global dissipation of the motion field v in the time interval $[0, 1]$ takes the form

$$\mathbf{Diss}\left[(v(t), \mathcal{O}(t))_{t\in[0,1]}\right] = \int_0^1 \int_{\mathcal{O}(t)} \mathbf{diss}[v]\, \mathrm{d}x\, \mathrm{d}t . \tag{2}$$

This is the same concept as employed by Dupuis et al. [23] and Miller et al. [53] in their pioneering diffeomorphism approach. They minimize a dissipation functional under the simplifying assumption that the material behaves equally viscous inside and outside the object. Also, $\mathbf{diss}[v] = \frac{\lambda}{2}(\mathrm{tr}\epsilon[v])^2 + \mu\mathrm{tr}(\epsilon[v]^2)$ is replaced by a higher-order quadratic form $Lv\cdot v$ which plays the role of the local rate of dissipation in a multipolar fluid model [57]. Multipolar fluids are characterized by the fact that the stresses depend on higher spatial derivatives of the velocity. If the quadratic form associated with L acts only on $\epsilon[v]$ and is symmetric, then rigid body motion invariance is incorporated in the multipolar fluid model (cf. section "Viscous Fluid-Based Shape Space"). In contrast to this approach, we here measure the rate of dissipation differently inside and outside the object and rely on classical (monopolar) material laws from fluid mechanics.

On this physical background, we will now derive a Riemannian structure on the space of shapes $\mathcal{S}$ in an admissible class of shapes $\mathbf{S}$. The associated metric $\mathcal{G}_{\mathcal{S}}$ on the (infinite-dimensional) manifold $\mathbf{S}$ is in abstract terms a bilinear mapping that assigns each element $\mathcal{S} \in \mathbf{S}$ an inner product on variations $\delta\mathcal{S}$ of $\mathcal{S}$ (cf. section "Recalling the Finite-Dimensional Case" above). The associated length of a tangent vector $\delta\mathcal{S}$ is given by $\|\delta\mathcal{S}\| = \sqrt{\mathcal{G}_{\mathcal{S}}(\delta\mathcal{S}, \delta\mathcal{S})}$. Furthermore, as we have already seen above, the length of a differentiable curve $\mathcal{S} : [0, 1] \to \mathbf{S}$ is then defined by $\mathbf{L}[\mathcal{S}] = \int_0^1 \|\dot{\mathcal{S}}(t)\|\, \mathrm{d}t = \int_0^1 \sqrt{\mathcal{G}_{\mathcal{S}(t)}\left(\dot{\mathcal{S}}(t), \dot{\mathcal{S}}(t)\right)}\, \mathrm{d}t$, where $\dot{\mathcal{S}}(t)$ is the temporal variation of $\mathcal{S}$ at time t. The Riemannian distance between two shapes $\mathcal{S}_A$ and $\mathcal{S}_B$ on $\mathbf{S}$ is given as the minimal length taken over all curves with $\mathcal{S}(0) = \mathcal{S}_A$ and $\mathcal{S}(1) = \mathcal{S}_B$ or equivalently (cf. section "Recalling the Finite-Dimensional Case" above) as the length of a minimizer of the functional $\int_0^1 \mathcal{G}_{\mathcal{S}(t)}\left(\dot{\mathcal{S}}(t), \dot{\mathcal{S}}(t)\right)\, \mathrm{d}t$. For shapes $\mathcal{S} \in \mathbf{S}$, an infinitesimal variation $\delta\mathcal{S}$ of a shape $\mathcal{S} = \partial\mathcal{O}$ is associated with

a transport field $v : \overline{\mathcal{O}} \to \mathbb{R}^d$. This transport field is obviously not unique. Indeed, given any vector field w on $\overline{\mathcal{O}}$ with $w(x) \in T_x\mathcal{S}$ for all $x \in \mathcal{S} = \partial\mathcal{O}$ (where $T_x\mathcal{S}$ denotes the $(d-1)$-dimensional tangent space to $\mathcal{S}$ at x), the transport field $v + w$ is another possible representation of the shape variation $\delta\mathcal{S}$. Let us denote by $\mathcal{V}(\delta\mathcal{S})$ the affine space of all these representations. As a geometric condition for $v \in \mathcal{V}(\delta\mathcal{S})$, we obtain $v(x) \cdot n[\mathcal{S}](x) = \delta\mathcal{S}(x) \cdot n[\mathcal{S}](x)$ for all $x \in \mathcal{S}$, where $n[\mathcal{S}](x) \in \mathbb{R}^d$ denotes the outer normal to $\mathcal{S} \subset \mathbb{R}^d$ in $x \in \mathcal{S}$. Given all possible representations, we are interested in the optimal transport, i.e., the transport leading to the least dissipation. Thus, using definition (1) of the local dissipation rate, we finally define the metric $\mathcal{G}_{\mathcal{S}}(\delta S, \delta S)$ as the minimal dissipation rate on motion fields v which are consistent with the variation of the shape δS,

$$\mathcal{G}_{\mathcal{S}}(\delta S, \delta S) := \min_{v \in \mathcal{V}(\delta\mathcal{S})} \int_{\mathcal{O}} \mathbf{diss}[v] \, \mathrm{d}x = \min_{v \in \mathcal{V}(\delta\mathcal{S})} \int_{\mathcal{O}} \frac{\lambda}{2} \left(\mathrm{tr}\epsilon[v]\right)^2 + \mu \, \mathrm{tr}\left(\epsilon[v]^2\right) \, \mathrm{d}x \, . \tag{3}$$

Let us remark that we distinguish explicitly between the metric $\mathbf{g}(v, v) := \int_{\mathcal{O}} \mathbf{diss}[v] \, \mathrm{d}x$ on motion fields and the metric $\mathcal{G}_{\mathcal{S}}(\delta\mathcal{S}, \delta\mathcal{S})$ on shape variations. Finally, integration in time leads to the total dissipation (2) to be invested in the transport along a path $(\mathcal{S}(t))_{t \in [0,1]}$ in the shape space $\mathbf{S}$. This implies the following definition of a time-continuous geodesic path in shape:

Definition 1 (Geodesic path). Given two shapes $\mathcal{S}_A$ and $\mathcal{S}_B$ in a shape space $\mathbf{S}$, a geodesic path between $\mathcal{S}_A$ and $\mathcal{S}_B$ is a curve $(\mathcal{S}(t))_{t \in [0,1]} \subset \mathbf{S}$ with $\mathcal{S}(0) = \mathcal{S}_A$ and $\mathcal{S}(1) = \mathcal{S}_B$ which is a local solution of

$$\min_{v(t) \in \mathcal{V}(\dot{\mathcal{S}}(t))} \mathbf{Diss}\left[(v(t), \mathcal{O}(t))_{t \in [0,1]}\right]$$

among all differentiable paths in $\mathbf{S}$.

The Riemannian distance between two shapes $\mathcal{S}_A$ and $\mathcal{S}_B$ induced by this definition is given by the length of the shortest (geodesic) path $\mathcal{S}(t)$ between the two shapes, i.e.,

$$d_{\text{viscous}}(\mathcal{S}_A, \mathcal{S}_B) = \mathbf{L}\left[(\mathcal{S}(t))_{t \in [0,1]}\right] \, .$$

Figure 3 shows two different paths between the same pair of shapes, one of them being a (numerically approximated) geodesic. Note that the chosen dissipation model combines the control of infinitesimal length changes via $\mathrm{tr}\left(\epsilon[v]^2\right)$, and the control of compression via $\mathrm{tr}\left(\epsilon[v]\right)^2$. Figure 4 evaluates the impact of these two terms on the shapes along a geodesic path.

State-Based, Path-Independent Elastic Setup

Now, objects bounded by a shape contour $\mathcal{S}$ are no longer composed of a viscous fluid but are considered to be elastic solids. To describe object deformations, we

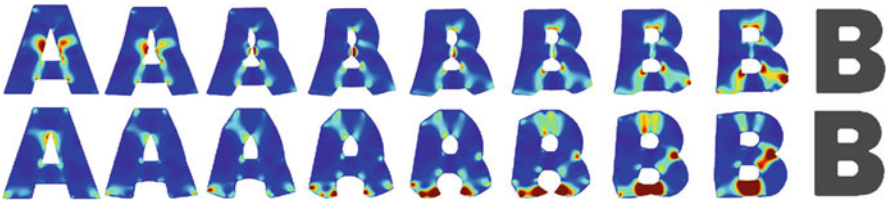


Fig. 3 A geodesic (*top*, path length $\mathbf{L} = 0.2225$ and total dissipation $\mathbf{Diss} = 0.0497$) and a non-geodesic path (*bottom*, $\mathbf{L} = 0.2886$, $\mathbf{Diss} = 0.0880$) between an A and a B. The intermediate shapes of the *bottom row* are obtained via linear interpolation between the signed distance functions of the end shapes. The local dissipation rate is color coded as

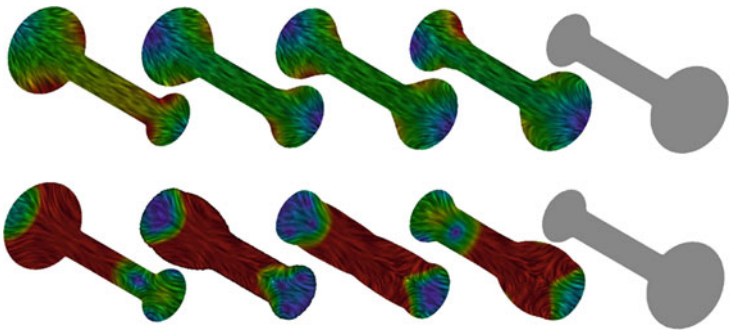


Fig. 4 Two geodesic paths between dumbbell shapes varying in the size of the ends. In the *top example*, the ratio λ/μ between the dissipation parameters is 0.01 (leading to rather independent compression and expansion of the ends since the associated change of volume implies relatively low dissipation), and 100 in the *bottom row* (now mass is actually transported from one end to the other). The underlying texture on the objects is aligned to the transport direction, and the absolute value of the velocity v is color coded as

aim for an elastic energy which is not restricted to small displacements and which is consistent with the first principles. Alongside the shape space modeling, we will recall some background from elasticity. For details, we refer to the comprehensive introductions in the books by Ciarlet [15] and Marsden and Hughes [47].

For two objects $\mathcal{O}_A$ and $\mathcal{O}_B$ with shapes $\mathcal{S}_A = \partial\mathcal{O}_A$ and $\mathcal{S}_B = \partial\mathcal{O}_B$, we assume a deformation ϕ to be defined on $\overline{\mathcal{O}}_A$ and constrained by the assumption $\phi(\mathcal{S}_A) = \mathcal{S}_B$. For practical reasons, one might consider $\mathcal{O}_A$ to be embedded in a very soft elastic material occupying $\Omega \setminus \mathcal{O}_A$ for some computational domain Ω. There is an elastic energy $\mathcal{W}_{\text{deform}}[\phi, \mathcal{O}_A]$ associated with the deformation $\phi : \Omega \to \mathbb{R}^d$. By definition, elastic means that this energy solely depends on the state and not on the path along which the deformation proceeds in time. More precisely, for so-called hyper-elastic materials, $\mathcal{W}_{\text{deform}}[\phi, \mathcal{O}_A]$ is the integral of an energy density W depending solely on the Jacobian $\mathcal{D}\phi$ of the deformation ϕ, i.e.,

$$\mathcal{W}_{\text{deform}}[\phi, \mathcal{O}_A] = \int_{\mathcal{O}_A} W(\mathcal{D}\phi)\, \mathrm{d}x\,. \tag{4}$$

This elastic energy is considered as a dissimilarity measure between the shapes $\mathcal{S}_A$ and $\mathcal{S}_B$. As a fundamental requirement, one postulates the invariance of the deformation energy with respect to rigid body motions, $\mathcal{W}_{\text{deform}}[Q \circ \phi + b, \mathcal{S}_A] = \mathcal{W}_{\text{deform}}[\phi, \mathcal{S}_A]$ for any orthogonal matrix $Q \in SO(d)$ and translation vector $b \in \mathbb{R}^d$ (the axiom of frame indifference in continuum mechanics). From this, one deduces that the energy density only depends on the right Cauchy–Green deformation tensor $\mathcal{D}\phi^{\mathrm{T}}\mathcal{D}\phi$. Hence, there is a function $\tilde{W} : \mathbb{R}^{d,d} \longrightarrow \mathbb{R}$ such that the energy density W satisfies $W(F) = \tilde{W}(F^{\mathrm{T}}F)$ for all $F \in \mathbb{R}^{d,d}$. The Cauchy–Green deformation tensor geometrically represents the metric measuring the deformed length in the undeformed reference configuration. For an isotropic material and for $d = 3$, the energy density W can be further rewritten as a function $\hat{W}(I_1, I_2, I_3)$ solely depending on the principal invariants of the Cauchy–Green tensor, namely, $I_1 = \mathrm{tr}(\mathcal{D}\phi^{\mathrm{T}}\mathcal{D}\phi)$, controlling the local average change of length; $I_2 = \mathrm{tr}\left(\mathrm{cof}(\mathcal{D}\phi^{\mathrm{T}}\mathcal{D}\phi)\right)$ ($\mathrm{cof}F := \det F\, F^{-T}$), reflecting the local average change of area; and $I_3 = \det(\mathcal{D}\phi^{\mathrm{T}}\mathcal{D}\phi)$, which controls the local change of volume. For a detailed discussion, we refer to [15,78]. We shall furthermore assume that the energy density is polyconvex [18], i.e., a convex function of $\mathcal{D}\phi$, $\mathrm{cof}\mathcal{D}\phi$, and $\det\mathcal{D}\phi$, and that isometries, i.e., deformations with $\mathcal{D}\phi^{\mathrm{T}}(x)\mathcal{D}\phi(x) = \mathbb{1}$, are local minimizers with $W(\mathcal{D}\phi) = \tilde{W}(\mathbb{1}) = 0$ [15]. Typical energy densities in this class are of the form

$$\hat{W}(I_1, I_2, I_3) = a_1 I_1^{\frac{p}{2}} + a_2 I_2^{\frac{q}{2}} + \Gamma(I_3) \tag{5}$$

for $a_1, a_2 > 0$ and a convex function $\Gamma : [0, \infty) \to \mathbb{R}$ with $\Gamma(I_3) \to \infty$ for $I_3 \to 0$ and $I_3 \to \infty$. In nonlinear elasticity, such material laws have been proposed by Ogden [58], and for $p = q = 2$ (the case considered in our computations), we obtain the Mooney–Rivlin model [15]. The built-in penalization of volume shrinkage, i.e., $\hat{W}(I_1, I_2, I_3) \overset{I_3 \to 0}{\longrightarrow} \infty$, enables us to control local injectivity (cf. [2]).

Incorporation of such a nonlinear elastic energy allows to describe large deformations with strong material and geometric nonlinearities, which cannot be treated by a linear elastic approach (cf. Hong et al. [36]). Furthermore, it balances in an intrinsic way expansion and collapse of the elastic objects and hence frees us from imposing artificial boundary conditions or constraints.

As in the previous section, the local force per area, induced by the deformation, is described at a point $\phi(x) \in \phi(\mathcal{O})$ by the Cauchy stress tensor σ. It is related to the first Piola–Kirchhoff stress tensor $\sigma^{\text{ref}} = W_{,F}(\mathcal{D}\phi) := \frac{\partial W(F)}{\partial F}|_{F=\mathcal{D}\phi}$, which measures the force density in the undeformed reference configuration, by $\sigma^{\text{ref}} = \sigma \circ \phi\, \mathrm{cof}\mathcal{D}\phi$.

Based on these concepts from nonlinear elasticity, we can now define a dissimilarity measure on shapes

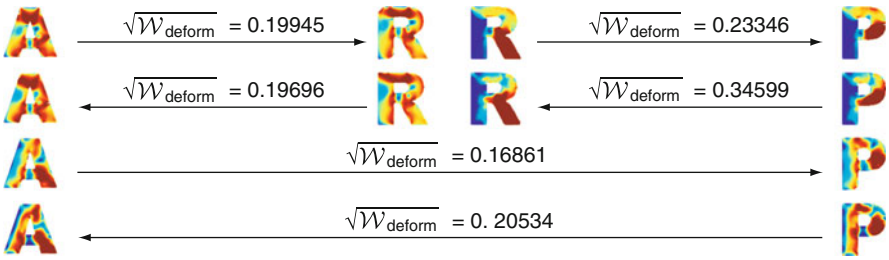


Fig. 5 Example of elastic dissimilarities between different shapes. The *arrows* indicate the direction of the deformation, the color coding represents the local deformation energy density (in the reference as well as the deformed state)

$$d_{\text{elast}}(\mathcal{S}_A, \mathcal{S}_B) := \min_{\phi, \phi(\mathcal{S}_A)=\mathcal{S}_B} \sqrt{\mathcal{W}_{\text{deform}}[\phi, \mathcal{O}_A]}\,. \tag{6}$$

Figure 5 shows some applications of this measure. Obviously, the elastic energy is in general not symmetric so that $d_{\text{elast}}(\mathcal{S}_A, \mathcal{S}_B) \neq d_{\text{elast}}(\mathcal{S}_B, \mathcal{S}_A)$. Indeed, by construction, $d_{\text{elast}}(\cdot, \cdot)$ does not impose a metric structure on the space of shapes (we refer to section "Conceptual Differences Between the Path- and State-Based Dissimilarity Measures" for a detailed discussion). Nevertheless, it can be applied to develop physically sound statistical tools for shapes such as shape averaging and a PCA on shapes, as outlined below in section "Elasticity-Based Shape Space."

Let us make a brief remark on the mathematical relation between the two different concepts of elasticity and viscous fluids. If we assume the Hessian of the energy density W at the identity to be given by $W_{,FF}(\mathbb{1})(G, G) = \lambda(\text{tr}G)^2 + \frac{\mu}{2}\text{tr}\left((G + G^T)^2\right)$ (which can be realized in (5) for a particular choice of a_1, a_2, and Γ, depending on the exponents p and q), then by the ansatz $\phi(x) = x + \tau v(x)$ and a second-order Taylor expansion, we obtain

$$\begin{aligned} W(\mathcal{D}\phi) &= W(\mathbb{1}) + \tau W_{,F}(\mathbb{1})(\mathcal{D}v) + \frac{\tau^2}{2} W_{,FF}(\mathbb{1})(\mathcal{D}v, \mathcal{D}v) + O(\tau^3) \\ &= 0 + 0 + \tau^2 \left(\frac{\lambda}{2} (\text{tr}\mathcal{D}v)^2 + \frac{\mu}{4} \text{tr}\left(\left(\mathcal{D}v + (\mathcal{D}v)^{\text{T}} \right)^2 \right) \right) + O(\tau^3)\,. \end{aligned} \tag{7}$$

In effect, the Hessian of the nonlinear elastic energy leads to the energy density in linearized, isotropic elasticity

$$W^{\text{lin}}(\mathcal{D}u) = \frac{\lambda}{2}\ (\text{tr}\epsilon[u])^2 + \mu\, \text{tr}\left(\epsilon[u]^2\right) \tag{8}$$

for displacements u with $\phi(x) = x + u(x)$. This energy density, acting on displacements u, formally coincides with the local dissipation rate $\mathbf{diss}[v]$, acting on velocity fields v, in the viscous flow approach.

Finally, let us deal with the hard constraint $\phi(\mathcal{S}_A) = \mathcal{S}_B$, which is often inadequate in applications. Due to local shape fluctuations or noise in the shape acquisition, the shape $\mathcal{S}_A$ frequently contains details that are not present in $\mathcal{S}_B$ and vice versa. These defects would imply high energies in a strict 1–1 matching approach. Hence, we have to relax the constraint and introduce some penalty functional. Here, we either measure the symmetric difference of the input shapes $\mathcal{S}_A$ and the pullback $\phi^{-1}(\mathcal{S}_B)$ of the shape $\mathcal{S}_B$ given by

$$\mathcal{F}[\mathcal{S}_A, \phi, \mathcal{S}_B] = \mathcal{H}^{d-1}\left(\mathcal{S}_A \triangle \phi^{-1}(\mathcal{S}_B)\right), \tag{9}$$

where $A \triangle B = A \setminus B \cup B \setminus A$, or alternatively the volume mismatch

$$\mathcal{F}[\mathcal{S}_A, \phi, \mathcal{S}_B] = \mathrm{vol}\left(\mathcal{O}_A \triangle \phi^{-1}(\mathcal{O}_B)\right). \tag{10}$$

Conceptual Differences Between the Path- and State-Based Dissimilarity Measures

The concept of the state-based, elastic approach to dissimilarity measurement between shapes differs significantly from the path-based viscous flow approach. In the elastic setup, the axiom of elasticity implies that the energy at the deformed configuration $\mathcal{S}_B = \phi(\mathcal{S}_A)$ is independent of the path from shape $\mathcal{S}_A$ to shape $\mathcal{S}_B$ along which the deformation is generated in time. Hence, there is no notion of the shortest paths if we consider a purely elastic shape model, and different from a path-based approach, there might not even exist an intermediate shape $\mathcal{S}_C$ with $d_{\text{elast}}(\mathcal{S}_A, \mathcal{S}_B) = d_{\text{elast}}(\mathcal{S}_A, \mathcal{S}_C) + d_{\text{elast}}(\mathcal{S}_C, \mathcal{S}_B)$.

Unlike in the elasticity model, in the Newtonian model of viscous fluids, the rate of dissipation and the induced stresses solely depend on the gradient of the motion field v. Even though the dissipation functional (2) looks like the deformation energy from linearized elasticity as outlined above, the underlying physics is only related in the sense that an infinitesimal displacement in the fluid leads to stresses caused by viscous friction, and these stresses are immediately absorbed via dissipation.

Surely, every (path-based) Riemannian space is metrizable (and in that sense state-based), and for many sufficiently regular (state-based) metric spaces, we can devise a corresponding (path-based) Riemannian metric. However, from our mechanical perspective, the conceptual difference between the path-based, viscous and the state-based elastic approach is striking. In the *path-based* approach, the structure of the space is too complicated for a closed formula of the geodesic distance, so that the actual computation of a path is required. In the *state-based* approach, there is either no underlying path (i.e., no $\mathcal{S}(t)_{t\in[0,1]}$ such that for any $0 \le t_1 \le t_2 \le t_3 \le 1$, we have $d(\mathcal{S}(t_1), \mathcal{S}(t_3)) = d\left(\mathcal{S}(t_1), \mathcal{S}(t_2)\right) + d\left(\mathcal{S}(t_2), \mathcal{S}(t_3)\right)$), or the shape space structure is simple enough to allow for a closed formula of the geodesic distance as in Euclidean space.

Mathematically, the path-based nature of the viscous flow approach and the fact that an inversion of the motion field $v \to -v$ leads to a path from shape $\mathcal{S}_B$ to $\mathcal{S}_A$ in shape space with the same dissipation and length, i.e.,

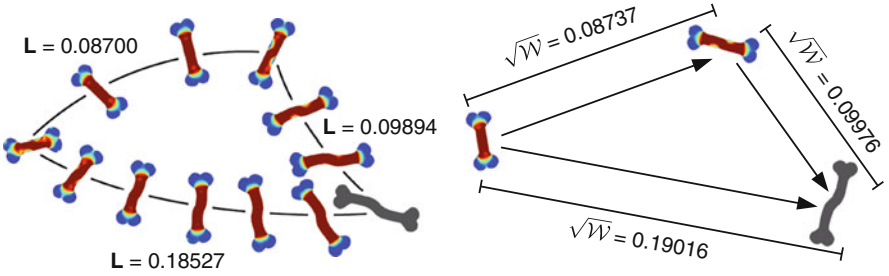


Fig. 6 *Left*: viscosity-based (time-discrete) geodesics between the shapes at the corners (the shapes are taken from [31]). The triangle inequality holds. *Right*: elastic dissimilarities $d_{\text{elast}}(\cdot,\cdot) = \sqrt{\mathcal{W}} \equiv \sqrt{\mathcal{W}_{\text{deform}}}$ between the same shapes, where the *arrows point* from the reference to the deformed configuration. The triangle inequality does not hold

Fig. 7 The state-based elastic dissimilarity measure d_{elast} is not symmetric (as opposed to the path-based, viscous distance d_{viscous}): In this example, it costs much more energy to drag out the protrusion than to push it in. The color coding represents the local deformation energy density in the reference and the deformed configuration

$$\mathbf{Diss}\left[(v(t),\mathcal{O}(t))_{t\in[0,1]}\right] = \mathbf{Diss}\left[(-v(1-t),\mathcal{O}(1-t))_{t\in[0,1]}\right]$$

ensure that the associated distance d_{viscous} is actually a metric. In particular, the symmetry condition $d_{\text{viscous}}(\mathcal{S}_A,\mathcal{S}_B) = d_{\text{viscous}}(\mathcal{S}_B,\mathcal{S}_A)$ and the triangle inequality $d_{\text{viscous}}(\mathcal{S}_A,\mathcal{S}_C) \le d_{\text{viscous}}(\mathcal{S}_A,\mathcal{S}_B) + d_{\text{viscous}}(\mathcal{S}_B,\mathcal{S}_C)$ hold. As we have already seen, the symmetry condition does not hold for the elastic dissimilarity measure. Also, the triangle inequality cannot be expected to hold. Indeed, if a deformation $\phi_{A,B}$ maps $\mathcal{O}_A$ onto $\mathcal{O}_B$ and a deformation $\phi_{B,C}$ maps $\mathcal{O}_B$ onto $\mathcal{O}_C$, then $\phi_{A,C} := \phi_{B,C} \circ \phi_{A,B}$ deforms $\mathcal{O}_A$ onto $\mathcal{O}_C$. However, based on our elastic model, $\mathcal{O}_B$ is considered to be stress free when applying the deformation $\phi_{B,C}$ (although it is actually obtained as the image of object $\mathcal{O}_A$ under the deformation $\phi_{A,B}$). Hence, the "history" of the deformation $\phi_{A,B}$ is lost when measuring the energy of $\phi_{B,C}$. In addition, the energy density is highly nonlinear. As a consequence, in general, we *c*annot expect $d_{\text{elast}}(\mathcal{S}_A,\mathcal{S}_C) \le d_{\text{elast}}(\mathcal{S}_A,\mathcal{S}_B) + d_{\text{elast}}(\mathcal{S}_B,\mathcal{S}_C)$. Indeed, Fig. 6 gives an example where the triangle inequality holds in the viscous, path-based and fails in the elastic, state-based approach. Furthermore, Fig. 7 depicts another example for the lack of symmetry already apparent in Fig. 5 with a particularly pronounced mechanical difference of the two dissimilarity measures.

4 Numerical Methods and Case Examples

Elasticity-Based Shape Space

In this section, we will perform a statistical analysis on shapes up to the second moment, i.e., we will consider shape averaging and a principal component analysis on shapes as two exemplary applications of the state-based elastic shape space.

Elastic Shape Averaging

As usual, we consider objects $\mathcal{O}$ as open sets in $\mathbb{R}^d$ with the object shape given as $\mathcal{S} := \partial\mathcal{O}$. Given n sufficiently regular shapes $\mathcal{S}_i = \partial\mathcal{O}_i$, $i = 1, \dots, n$, we are interested in an average shape which reflects the geometric characteristics of the input shapes in a physically intuitive manner. Suppose $\mathcal{S} = \partial\mathcal{O} \subset \mathbb{R}^d$ denotes a candidate for this unknown shape. As it is characteristic for the elastic approach, the similarity of the input shapes $\mathcal{S}_i$ to $\mathcal{S}$ is measured by taking into account optimal elastic deformations $\phi_i : \overline{\mathcal{O}}_i \to \mathbb{R}^d$ with $\phi_i(\mathcal{S}_i) = \mathcal{S}$. The elastic energy $\mathcal{W}_{\text{deform}}[\phi_i, \mathcal{O}_i]$ of these deformations has the interpretation of a dissimilarity measure (cf. section "State-Based, Path-Independent Elastic Setup"), so that we obtain a natural definition of an average shape as the minimizer of the sum of these terms (cf. Sect. 2).

Definition 2 (Elastic shape average). Given shapes $\mathcal{S}_1, \dots, \mathcal{S}_n$ in some shape space $\mathbf{S}$, the elastic shape average $\mathcal{S}$ is the minimizer of

$$\sum_{i=1}^{n} d_{\text{elast}}(\mathcal{S}_i, \mathcal{S})^2 = \sum_{i=1}^{n} \inf_{\phi_i : \overline{\mathcal{O}}_i \to \mathbb{R}^d, \phi_i(\mathcal{S}_i) = \mathcal{S}} \mathcal{W}_{\text{deform}}[\phi_i, \mathcal{O}_i].$$

If the input objects $\mathcal{O}_i$ have Lipschitz boundary and the integrand of the deformation energy $\mathcal{W}_{\text{deform}}[\phi_i, \mathcal{O}_i] = \int_{\mathcal{O}_i} W(\mathcal{D}\phi_i)\,\mathrm{d}x$ is polyconvex and bounded below by $C_1\|\mathcal{D}\phi_i\|^p - C_2$ for $p > d$, $C_1, C_2 > 0$, the existence of a Hölder-continuous elastic shape average and deformations $\phi_i \in W^{1,p}(\mathcal{O}_i)$ which realize the above infimum is guaranteed [81].

An example of a shape average is provided in Fig. 8. Obviously, the process of shape averaging is a constrained variational problem in which we simultaneously have to minimize over n deformations ϕ_i and the unknown shape $\mathcal{S}$ under the n constraints $\phi_i(\mathcal{S}_i) = \mathcal{S}$.

The necessary conditions for a set of minimizing deformations are the corresponding Euler–Lagrange equations. As usual, inner variations of one of the deformations lead to the classical system of PDEs $\operatorname{div} W_{,F}(\mathcal{D}\phi_i) = 0$ for every deformation ϕ_i on $\mathcal{O}_i \setminus \mathcal{S}_i$, meaning a divergence-free, equilibrized stress field (cf. section "State-Based, Path-Independent Elastic Setup"). Furthermore, the coupling between the deformations via the constraints $(\phi_i(\mathcal{S}_i) = \mathcal{S})_{i=1,\dots,n}$ allows to derive a stress balance relation on $\mathcal{S}$: Consistent variation of all deformations ϕ_i and the average $\mathcal{S}$ by some displacement $u : \overline{\mathcal{O}} \to \mathbb{R}^d$ via $(\mathbb{1} + \delta u) \circ \phi_i$ and $(\mathbb{1} + \delta u)(\mathcal{S})$

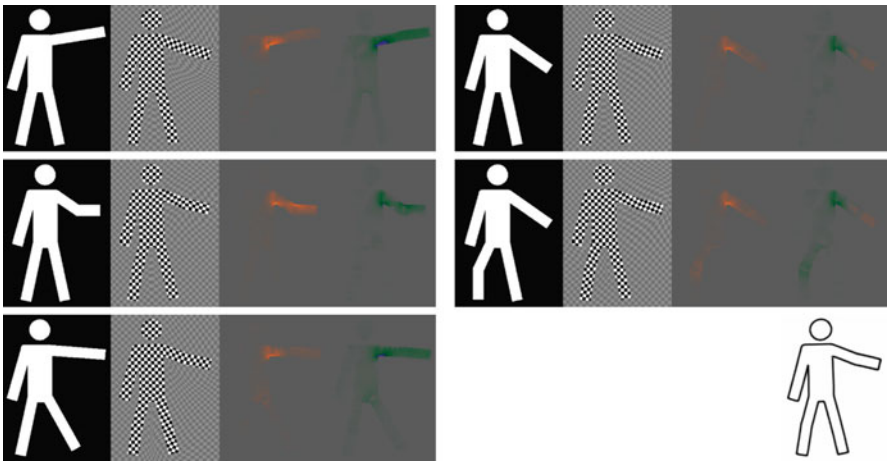


Fig. 8 Elastic shape average (*bottom right*) of five human silhouettes. For the computation, all shapes have actually been described as phase fields, and the elastic deformations are extended outside the input objects $\mathcal{O}_i$ (cf. section "Shapes Described via Phase Fields"). The objects $\mathcal{O}_i$ are depicted along with their deformations ϕ_i (acting on a checkerboard) and the distribution of local length change $\frac{1}{\sqrt{2}}\|\mathcal{D}\phi_i\|$ and volume change $\det(\mathcal{D}\phi_i)$ (range $[0.97, 1.03]$ color coded as)

results in the optimality condition $\frac{d}{d\delta}\sum_{i=1}^{n}\mathcal{W}_{\text{deform}}[(\mathbb{1}+\delta u)\circ\phi_i, \mathcal{O}_i]\big|_{\delta=0} = 0$, which after integration by parts leads to $\sum_{i=1}^{n}\int_{\mathcal{S}_i} W_{,F}(\mathcal{D}\phi_i)(u\circ\phi_i)\cdot\nu[\mathcal{S}_i]\,\mathrm{d}a[\mathcal{S}_i] = 0$ for the outer normal $\nu[\mathcal{S}_i]$ to $\mathcal{S}_i$. We have here exploited $\operatorname{div} W_{,F}(\mathcal{D}\phi_i) = 0$ on $\mathcal{O}_i \setminus \mathcal{S}_i$. Now, we consider displacements u with local support and let this support collapse at some point x on $\mathcal{S}$. This yields the pointwise condition

$$0 = \sum_{i=1}^{n}\left(\sigma_i^{\text{ref}}\nu[\mathcal{S}_i]\,\mathrm{d}a[\mathcal{S}_i]\right)\left(\phi_i^{-1}(x)\right) \quad \text{and thus} \quad 0 = \sum_{i=1}^{n}(\sigma_i\,\nu[\mathcal{S}])(x) \tag{11}$$

for $x \in \mathcal{S}$, where we have used the relation

$$\left(\sigma_i^{\text{ref}}\nu[\mathcal{S}_i]\,\mathrm{d}a[\mathcal{S}_i]\right)\left(\phi_i^{-1}(x)\right) = (\sigma_i\,\nu[\mathcal{S}]\,\mathrm{d}a[\mathcal{S}])(x)$$

between the first Piola–Kirchhoff stress $\sigma_i^{\text{ref}} = W_{,F}(\mathcal{D}\phi_i)$ and Cauchy stress $\sigma_i = \left(\sigma_i^{\text{ref}}(\operatorname{cof}\mathcal{D}\phi_i)^{-1}\right)\circ\phi_i^{-1}$. Hence, the shape average can be interpreted as that stable shape at which the boundary stresses of all deformed input shapes balance each other (Fig. 9). Obviously, there is a straightforward generalization involving jumps of normal stresses on interior interfaces in case of multicomponent objects.

In order to ensure a certain regularity of the average shape $\mathcal{S}$, in addition to the sum of deformation energies in Definition 2, one can consider a further energy contribution which acts as a prior on $\mathcal{S}$ in the variational approach. In the exemplary

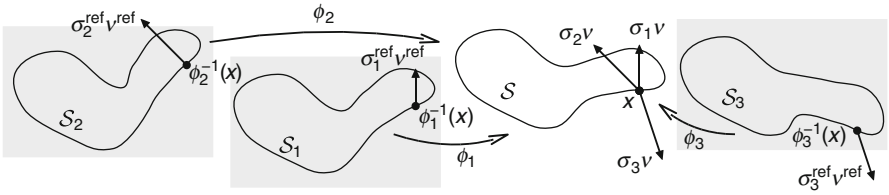


Fig. 9 Sketch of the pointwise stress balance relation on the averaged shape

Fig. 10 Average of 18 hand silhouettes (Taken from [16])

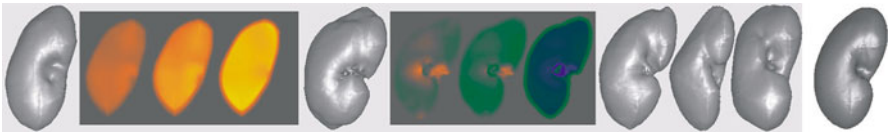


Fig. 11 Five segmented kidneys and their average (*right*). For the first two input kidneys, the distribution of $\frac{1}{\sqrt{3}}\|\mathcal{D}\phi_i\|$, $\frac{1}{\sqrt{3}}\|\mathrm{cof}(\mathcal{D}\phi_i)\|$, and $\det(\mathcal{D}\phi_i)$ is shown on sagittal cross sections (the range $[0.85, 1.15]$ is color coded as). While the first kidney is dilated toward the average, the second is compressed

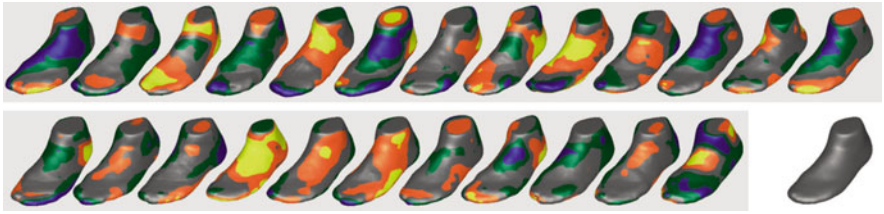


Fig. 12 Twenty-four given foot shapes (Courtesy of Adidas), textured with the distance to the surface of the average foot (*bottom right*). Values range from 6 mm inside the average foot to 6 mm outside, color coded as

computations shown (Figs. 10–12), the $(d-1)$-dimensional Hausdorff measure $\mathcal{L}[\mathcal{S}] = \mathcal{H}^{d-1}(\mathcal{S})$ has been employed as regularization.

Elasticity-Based PCA

As already explained in section "Recalling the Finite-Dimensional Case," a principal component analysis (PCA) is a linear statistical tool which decomposes a vector space into the direct sum of orthogonal subspaces. These subspaces are ordered according to the strength of variation which occurs along each subspace within a random set of sample vectors. We would like to interpret a given set of input shapes $\mathcal{S}_1, \ldots, \mathcal{S}_n$ as such a random sample and perform a corresponding PCA; however,

due to the linearity of a PCA, we first have to identify linear representatives for each shape on which a PCA can then be performed. For a Riemannian shape space, we have outlined in section "Recalling the Finite-Dimensional Case" that such linear representatives are given by the logarithmic map of the input shapes, but we have also learned in section "Conceptual Differences Between the Path- and State-Based Dissimilarity Measures" that a state-based elastic shape space is incompatible with a Riemannian structure.

To prepare the definition of appropriate linear representatives of shapes in an elastic shape space, let us briefly review the physical concept of boundary stresses. By the Cauchy stress principle, each deformation $\phi_k : \mathcal{O}_k \to \mathcal{O}$ is characterized by pointwise boundary stresses on $\mathcal{S} = \partial\mathcal{O}$ in the deformed configuration. The stress at some point x on $\mathcal{S}$ is given by the application of the Cauchy stress tensor σ_k to the outer normal ν on $\mathcal{S}$. The resulting stress $\sigma_k \nu$ is a force density acting on a local surface element of $\mathcal{S}$. The shape $\mathcal{S}$ is in an equilibrium configuration if the opposite force is applied as an external surface load (cf. Fig. 9). Otherwise, by the axiom of elasticity, releasing the object $\mathcal{O}$, the elastic body will snap back to the original reference configuration $\mathcal{O}_k$. Let us assume the relation between the energetically favorable deformation and its induced stresses to be one to one, so that the average shape $\mathcal{S}$ can be described in terms of the input shape $\mathcal{S}_k$ and the boundary stress $\sigma_k \nu$, and we write $\mathcal{S} = \mathcal{S}_k[\sigma_k \nu]$. Upon scaling the stress with a weight $t \in [0, 1]$, we obtain a one-parameter family of shapes $\mathcal{S}(t) = \mathcal{S}_k[t\sigma_k \nu]$, connecting $\mathcal{S}_k = \mathcal{S}(0)$ with $\mathcal{S} = \mathcal{S}(1)$. Thus, we can regard $\sigma_k \nu$ as a representative of shape $\mathcal{S}_k$ in the linear space of vector fields on $\mathcal{S}$.

Physically, it is more intuitive to identify a displacement u_k instead of the normal stress $\sigma_k \nu$ as the representative of an input shape $\mathcal{S}_k$. Hence, let us study how the average shape $\mathcal{S}$ varies if we increase the impact of a particular input shape $\mathcal{S}_k$ for some $k \in \{1, \ldots, n\}$. For this purpose, we apply the Cauchy stress $\sigma_k \nu$ to the average shape $\mathcal{S}$, scaled with a small constant δ. This additional boundary stress $\delta\sigma_k \nu$ may be seen as a first Piola–Kirchhoff stress acting on the (reference) configuration $\mathcal{S}$. The elastic response is given by a correspondingly scaled displacement $u_k : \mathcal{O} \to \mathbb{R}^d$. Here, to properly incorporate the nonlinear nature of the second moment analysis, $\mathcal{O}$ should be interpreted as the compound object which is composed of all deformed and thus prestressed input objects $\phi_i(\mathcal{O}_i)$. This interpretation is reflected by the elastic material law employed to compute the displacements u_k. In detail, u_k is obtained as the minimizer of the free mechanical energy

$$\mathcal{E}_k[\delta, u] = \frac{1}{n} \sum_{i=1}^{n} \mathcal{W}_{\text{deform}}\big[(\mathbb{1} + \delta u) \circ \phi_i, \mathcal{O}_i\big] - \delta^2 \int_{\mathcal{S}} \sigma_k \nu \cdot u \, \mathrm{d}a \tag{12}$$

under the constraints $\int_{\mathcal{O}} u_k \, \mathrm{d}x = 0$ and $\int_{\mathcal{O}} x \times u_k \, \mathrm{d}x = 0$ of zero average translation and rotation. These displacements u_k are considered as representatives of the variation of the average shape $\mathcal{S}$ with respect to the input shape $\mathcal{S}_k$, on which a PCA will be performed.

As long as $F \mapsto W(F)$ is not quadratic in F, u_k still solves a nonlinear elastic problem. The advantage of this nonlinear variational formulation is that it is of the same type as the one for shape averaging, and it encodes in a natural way the compound elasticity configuration of the averaged shape domain $\mathcal{O}$. However, for the linearization of shape variations, we are actually only interested in the displacements δu_k for small δ. Therefore, we consider the limit of the Euler–Lagrange equations for $\delta \to 0$ and, after a little algebra, obtain u_k as the solution of the linearized elasticity problem

$$\operatorname{div}\big(\mathbf{C}\,\epsilon[u]\big) = 0 \text{ in } \mathcal{O}\,, \quad \mathbf{C}\,\epsilon[u]\,\nu = \sigma_k \nu \text{ on } \mathcal{S} \tag{13}$$

for the symmetrized displacement gradient $\epsilon[u] = \frac{1}{2}(\mathcal{D}u + \mathcal{D}u^T)$ under the constraints $\int_{\mathcal{O}} u \,\mathrm{d}x = 0$ and $\int_{\mathcal{O}} x \times u \,\mathrm{d}x = 0$, where the in general inhomogeneous and anisotropic elasticity tensor $\mathbf{C}$ reads

$$\mathbf{C} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{\det \mathcal{D}\phi_i}\mathcal{D}\phi_i\, W_{,FF}(\mathcal{D}\phi_i)\mathcal{D}\phi_i^{\mathrm{T}}\right)\circ \phi_i^{-1}\,.$$

Next, for a PCA on the linearized shape variations u_k, we select a suitable inner product (metric) $g(u, \tilde{u})$ on displacements $u, \tilde{u} : \mathcal{O} \to \mathbb{R}^d$. Note that g induces a metric $\tilde{g}(\sigma\nu, \tilde{\sigma}\nu) := g(u, \tilde{u})$ on the associated boundary stresses so that instead of analyzing the u_k, the covariance analysis can equivalently be performed directly on the boundary stresses $\sigma_1\nu, \ldots, \sigma_n\nu$, which we originally derived as linear shape representatives. Indeed, the solvability condition $\int_{\mathcal{O}} \operatorname{div}(\mathbf{C}\nabla u)\,\mathrm{d}x = \int_{\mathcal{S}} \mathbf{C}\nabla u\nu \,\mathrm{d}a[\mathcal{S}]$ is fulfilled, and thus the solution u_k for given boundary stress $\sigma_k\nu = \mathbf{C}\nabla u\nu$ is uniquely determined up to a linearized rigid body motion (i.e., an affine displacement with skew-symmetric matrix representation), which is fixed by the conditions of zero mean displacement and angular momentum for u. Then, due to the linearity of the operator $\sigma\nu \mapsto u$, the metric $\tilde{g}$ is bilinear and symmetric as well, and its positive definiteness follows from the positive definiteness of g and the injectivity of the map $\sigma\nu \mapsto u$.

We consider two different inner products on displacements $u : \mathcal{O} \to \mathbb{R}^d$:

- *The L^2-product.* Given two square integrable displacements $u,\ \tilde{u}$, we define

$$g(u, \tilde{u}) := \int_{\mathcal{O}} u \cdot \tilde{u}\,\mathrm{d}x.$$

  This product weights local displacements equally on the whole object $\mathcal{O}$.
- *The Hessian of the energy as inner product.* Different from the L^2-metric, we now measure displacement gradients in a nonhomogeneous way. We define

$$g(u, \tilde{u}) := \int_{\mathcal{O}} \mathbf{C}\epsilon[u] : \epsilon[\tilde{u}]\,\mathrm{d}x$$

for displacements u, $\tilde{u}$ with square integrable gradients. Hence, the contribution to the inner product is larger in areas of the compound object which are in a significantly stressed configuration.

Given an inner product, we can define the covariance operator **Cov** by

$$\mathbf{Cov}\, u := \frac{1}{n} \sum_{k=1}^{n} g(u, u_k) u_k$$

(note that the stresses $\sigma_k \nu$ and thus also the displacements u_k have zero mean due to (11)). Obviously, **Cov** is symmetric positive definite on $\operatorname{span}(u_1, \ldots, u_n)$. Hence, we can diagonalize **Cov** on this finite-dimensional space and obtain a set of g-orthonormal eigenfunctions $w_k : \mathcal{O} \to \mathbb{R}^d$ and eigenvalues $\lambda_k > 0$ with $\mathbf{Cov}\, w_k = \lambda_k w_k$. These eigenfunctions can be considered as the principal modes of variation of the average object $\mathcal{O}$ and hence of the average shape $\mathcal{S}$, given the n sample shapes $\mathcal{S}_1, \ldots, \mathcal{S}_n$. Their eigenvalues encode the variation strength. The diagonalization of **Cov** can be performed by diagonalizing the symmetric matrix $\frac{1}{n}\big(g(u_i, u_j)\big)_{ij} = O \Lambda O^{\mathrm{T}}$, where $\Lambda = \operatorname{diag}(\lambda_1, \lambda_2, \ldots)$ and O is orthogonal. The eigenfunctions are then obtained as $w_k = \frac{1}{\sqrt{\lambda_k}} \sum_{j=1}^{n} O_{jk} u_j$.

Being displacements on $\mathcal{O}$, the modes of variation w_k can easily be visualized via a scalar modulation δw_k for varying δ (cf. the vizualization in Figs. 16–18 or the red lines in Figs. 13 and 15). If an amplified visualization of the modes is required, it is preferable to depict displacements w_δ^k which are defined as minimizers of the nonlinear variational energy $\frac{1}{n} \sum_{i=1}^{n} \mathcal{W}_{\text{deform}}\big[(\mathbb{1} + w) \circ \phi_i, \mathcal{O}_i\big] - \delta^2 \int_{\mathcal{S}} \mathbf{C} \nabla w_k \nu \cdot w \, \mathrm{d}a$ (cf. (12)).

Let us underline that this covariance analysis properly takes into account the usually strong geometric nonlinearity in shape analysis via the transfer of geometric shape variation to elastic stresses on the average shape, based on paradigms from nonlinear elasticity. Displacements or stresses are interpreted as the proper linearization of shapes. In abstract terms, either the space of displacements or stresses can be considered as the tangent space of shape space at the average shape, where the identification of displacements and stresses via (13) provides a suitable physical interpretation of stresses as shape variations.

The impact of the chosen metric. Naturally, the modes of variation depend on the chosen inner product. We have already mentioned that in order to be physically meaningful, the inner product should act on displacements u_k of the compound object (which is composed of all deformed input shapes). If instead the u_k were obtained by applying the boundary stresses $\sigma_k \nu$ to an object which just looks like the average shape but does not contain the information how strongly the input shapes had to be deformed to arrive at the average, we obtain a different result (Fig. 13, *left*): If the prestressed state of some object regions is neglected, it becomes easier to deform them which causes the prediction of stronger variations. Figure 13 also hints at the differences between the employed metrics: The L^2-metric pronounces

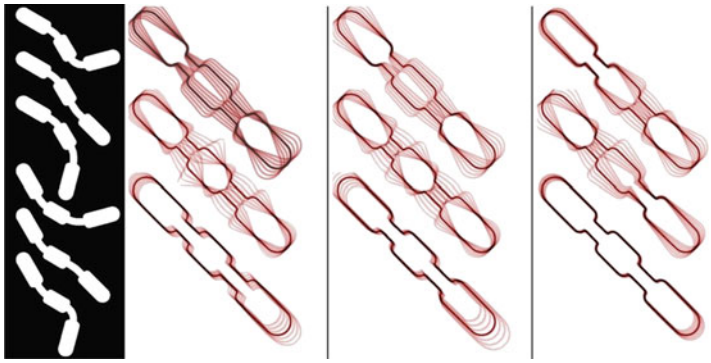


Fig. 13 First three dominant modes of variation for six input shapes (*left*), based on different metrics. *Left*: L^2-metric on displacements of a non-prestressed object (modes w_k with ratios $\frac{\lambda_k}{\lambda_1}$ of 1, 0.23, 0.07). *Middle*: L^2-metric on displacements of the compound object $\left(\frac{\lambda_k}{\lambda_1} = 1, 0.28, 0.03\right)$. *Right*: energy Hessian-based metric on displacements of the compound object $\left(\frac{\lambda_k}{\lambda_1} = 1, 0.61, 0.24\right)$

shape variations with large displacements even though they are energetically cheap (e.g., a rotation of some structure around a joint), while the Hessian of the elastic energy measures distances between displacements solely based on the associated change of elastic energy. Thus, displacements are weighted strongly in regions and directions which are significantly loaded.

The impact of the nonlinear elasticity model. Likewise, the particular choice of the nonlinear elastic energy density has a considerable effect on the average shape and its modes of variation. Figure 14 has been obtained using $W(\mathcal{D}\phi) = \frac{\mu}{2}\|\mathcal{D}\phi\|^2 + \frac{\lambda}{4}\det\mathcal{D}\phi^2 - \left(\mu + \frac{\lambda}{2}\right)\log\det\mathcal{D}\phi - \mu - \frac{\lambda}{4}$, where μ and λ are the coefficients of length and volume change penalization, respectively. A low penalization of volume changes apparently leads to independent compression and inflation at the dumbbell ends (left), while for deformations with a strong volume change penalization (right), material is squeezed from one end to the other. Here, the underlying metric is the based on the Hessian of the energy.

Figures 15–17 show the dominant modes of variation for the examples from the previous section. A statistical analysis of the hand shapes in Fig. 15 has also been performed in [16] and [28], where the shapes are represented as vectors of landmark positions. The average and the modes of variation are quite similar, representing different kinds of spreading the fingers. The dominant modes of variation for a set of 48 three-dimensional kidney shapes is depicted in Fig. 16, where for all modes w_k, we show the average (middle) and its variation according to δw_k for varying δ. Local structures seem to be quite well represented and preserved during the averaging process and the subsequent covariance analysis compared to, e.g., the PCA on kidney shapes in [26] where a medial representation is used.

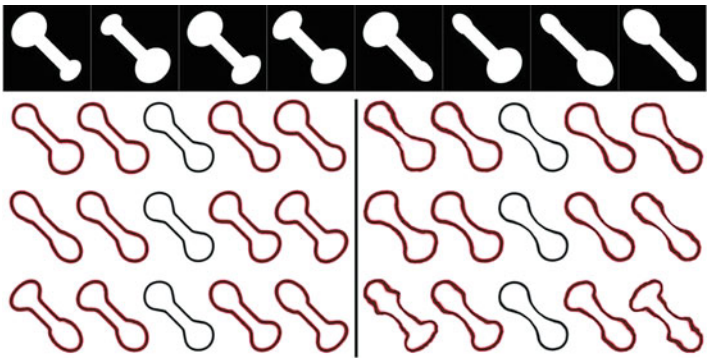


Fig. 14 First three modes of variation for eight dumbbell shapes, *left* for a 100 times stronger penalization of length than of volume changes (with ratios $\frac{\lambda_i}{\lambda_1}$ of 1, 0.22, 0.05), *right* for the reverse $\left(\frac{\lambda_i}{\lambda_1} = 1, 0.41, 0.07\right)$. Each row represents the variation of the average (*middle shape*) by δw_k for the mode w_k and varying δ

Fig. 15 First four modes of variation with ratios $\frac{\lambda_i}{\lambda_1}$ of 1, 0.88, 0.42, and 0.25 for the 18 hand silhouettes from Fig. 10

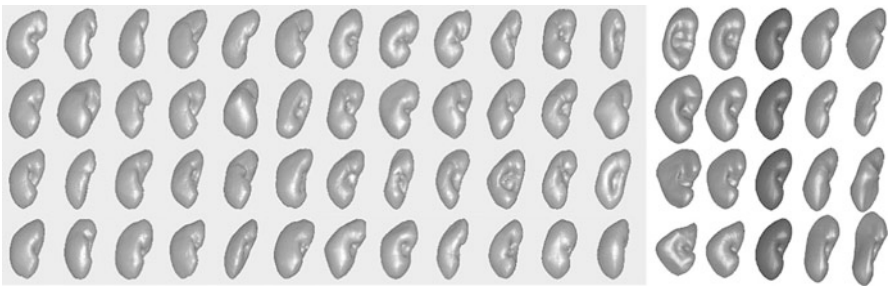


Fig. 16 Forty-eight input kidneys (Courtesy of Werner Bautz, radiology department at the University Hospital Erlangen, Germany) and their first four modes of variation with ratios $\frac{\lambda_i}{\lambda_1}$ of 1, 0.72, 0.37, and 0.31

The PCA of the 24 ft shapes from Fig. 12 is shown in Fig. 17 and is much more intuitive than the color coding in Fig. 12. The first mode apparently represents changing foot lengths, the second and third mode belong to different variants of combined width and length variation, and the fourth to sixth mode correspond to variations in relative heel position, ankle thickness, and instep height. Finally, Fig. 18 shows that the approach also works for image morphologies instead of

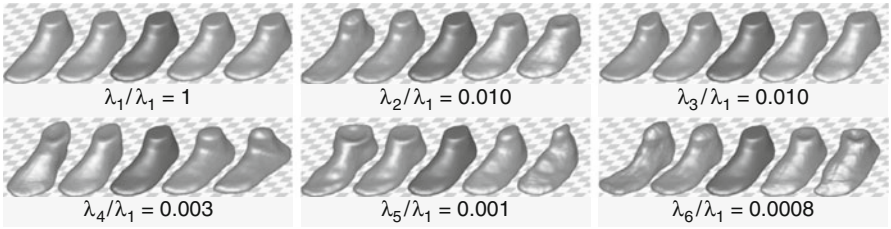


Fig. 17 The first six dominant modes of variation for the feet from Fig. 12

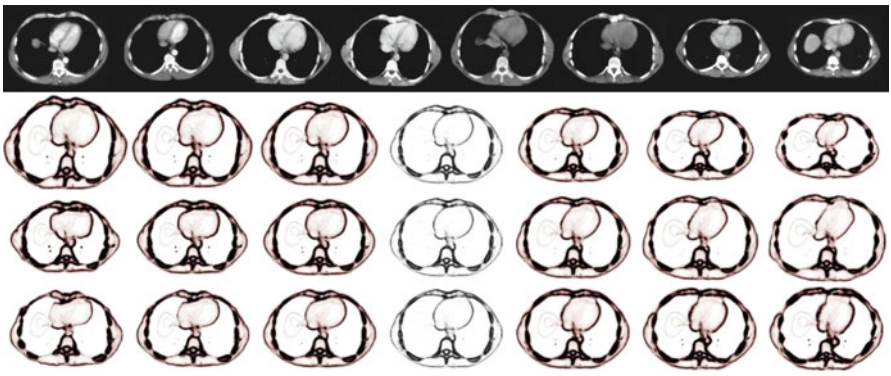


Fig. 18 Eight thorax CT scans from different patients (Courtesy of Bruno Wirth, urology department at the Hospital zum hl. Geist, Kempen, Germany) and their first three modes of variation with ratios $\frac{\lambda_i}{\lambda_1}$ of 1, 0.12, and 0.07. Note that the thin lines which can be seen left of the heart correspond to contours of the liver, which are only visible in the first and last input image

shapes, using thorax CT scans as input. Here, the image edge set is considered as the corresponding shape, which is typically quite complex and characterized by nested contours. The first mode of variation represents a variation in chest size, the next mode corresponds to a change of heart and scapula shape, while the third mode mostly concerns the rib position.

Viscous Fluid-Based Shape Space

As explained in section "Path-Based, Viscous Riemannian Setup," the viscous fluid shape space is by construction a (infinite-dimensional) Riemannian manifold and as such is based on the computation of shape paths as opposed to state-based approaches like the elastic shape space from the previous section. In the elastic, state-based approach, we have to find for each pair of shapes $\mathcal{S}_A = \partial\mathcal{O}_A$ and $\mathcal{S}_B = \partial\mathcal{O}_B$ one single optimal matching deformation $\phi : \mathcal{O}_A \to \mathbb{R}^d$ via which the similarity between $\mathcal{S}_A$ and $\mathcal{S}_B$ is determined. In contrast, here we require more information to measure the distance between the two shapes, namely, an optimal

velocity field $v(t) : \mathcal{O}(t) \to \mathbb{R}^d$ at each time t within the given time interval $[0, 1]$. In effect, this implies an increase of the dimension of the variational problem by the time component.

The two qualitatively different types of coordinates, the space coordinates (that span the space in which the shapes lie) and the time coordinate, are intuitively treated in different ways. One possibility is to regard the variational problem of computing a geodesic as a classical elliptic boundary value problem in time, in which each shape on a path seeks to be in equilibrium with its local neighborhood on the path. The equilibrizing force can be interpreted as an acceleration acting on the velocity field v. In this setting, it seems most natural to discretize first the time variable and approximate geodesics in shape space as discrete sequences $\mathcal{S}_0, \ldots, \mathcal{S}_K$ of shapes, where each shape is connected to and equilibrates with its neighbors and the path length along the discrete path $\mathcal{S}_0, \ldots, \mathcal{S}_K$ is approximated as a sum $\sum_{k=1}^{K} \tilde{d}(\mathcal{S}_{k-1}, \mathcal{S}_k)$ of approximations $\tilde{d}(\mathcal{S}_{k-1}, \mathcal{S}_k)$ of the geodesic distance between neighboring shapes. The distance $\tilde{d}$ can be based on a matching deformation energy which will be elaborated on further down.

An alternative view starts from the underlying velocity field which generates the geodesic. Dupuis et al. [23] and Beg et al. [3] consider shapes (or rather images) embedded in a domain $\Omega \subset \mathbb{R}^d$. These shapes deform according to smooth, compactly supported velocity fields $v \in L^2\left([0,1]; W_0^{n,2}(\Omega; \mathbb{R}^d)\right)$ with $n > 2 + \frac{2}{d}$. The regularity of the velocity fields is ensured by defining the path dissipation as $\int_0^1 \int_\Omega Lv \cdot v \, \mathrm{d}x \, \mathrm{d}t$ and the path length as $\int_0^1 \sqrt{\int_\Omega Lv \cdot v \, \mathrm{d}x} \, \mathrm{d}t$ for a differential operator L of sufficiently high order (cf. section "Path-Based, Viscous Riemannian Setup"). The corresponding shape deformation ϕ which is induced by the velocity field is obtained as the solution $\phi = \phi_1$ of the pointwise, Lagrangian ordinary differential equation $\frac{\mathrm{d}}{\mathrm{d}t}\phi_t(x) = v(\phi_t(x), t)$.

In the first approach, the computation of a geodesic was seen as the concatenation of a number of local subproblems, each of which represents the approximation of a geodesic segment between two intermediate shapes and each of which thus inherits the constraint that one shape is transferred exactly into the other. In contrast, in the second approach, we have one single constraint, acting at the end of the geodesic and expressing that the accumulated flow ϕ deforms the starting shape $\mathcal{S}_A$ into the final shape $\mathcal{S}_B$, $\phi(\mathcal{S}_A) = \mathcal{S}_B$.

Let us now focus on the first approach in which a geodesic path will be approximated via a finite sequence of shapes $\mathcal{S}_0, \ldots, \mathcal{S}_K$, connected by deformations $\phi_k : \mathcal{O}_{k-1} \to \mathbb{R}^d$ which are optimal in a variational sense and fulfil the constraint $\phi_k(\mathcal{S}_{k-1}) = \mathcal{S}_k$.

Given two shapes $\mathcal{S}_A$, $\mathcal{S}_B$ in some given space of shapes $\mathbf{S}$, we define a discrete path of shapes as a sequence of shapes $\mathcal{S}_0, \ldots, \mathcal{S}_K \in \mathbf{S}$ with $\mathcal{S}_0 = \mathcal{S}_A$ and $\mathcal{S}_K = \mathcal{S}_B$. For the time step $\tau = \frac{1}{K}$, the shape $\mathcal{S}_k$ is supposed to be an approximation of $\mathcal{S}(t_k)$ with $t_k = k\tau$, where $(\mathcal{S}(t))_{t\in[0,1]}$ is a continuous path connecting $\mathcal{S}_A = \mathcal{S}(0)$ and $\mathcal{S}_B = \mathcal{S}(1)$. For each pair of consecutive shapes $\mathcal{S}_{k-1}$ and $\mathcal{S}_k$, we now consider a matching deformation $\phi_k : \mathcal{O}_{k-1} \to \mathbb{R}^d$ which

satisfies $\phi_k(\mathcal{S}_{k-1}) = \mathcal{S}_k$. With each deformation ϕ_k, we associate a deformation energy $\mathcal{W}_{\text{deform}}[\phi_k, \mathcal{O}_{k-1}] = \int_{\mathcal{O}_{k-1}} W(\mathcal{D}\phi_k)\,\mathrm{d}x$ of the same type as described in section "State-Based, Path-Independent Elastic Setup." If appropriately chosen, this energy will ensure sufficient regularity and a 1–1 matching property for deformations ϕ_k with finite energy. As in elasticity, the energy is assumed to depend only on the local deformation, reflected by the Jacobian $\mathcal{D}\phi$. Yet, different from elasticity, we suppose the material to relax instantaneously so that object $\mathcal{O}_k$ is again in a stress-free configuration when applying ϕ_{k+1} at the next time step. Let us also emphasize that the stored energy does not depend on the deformation history as in most plasticity models in engineering. This energy is now employed to define time-discrete counterparts to the dissipation and length of continuous paths from section "Path-Based, Viscous Riemannian Setup."

Definition 3 (Discrete dissipation and discrete path length). Given a discrete path $\mathcal{S}_0, \ldots, \mathcal{S}_K \in \mathbf{S}$, its dissipation is defined as

$$\mathbf{Diss}_\tau(\mathcal{S}_0, \ldots, \mathcal{S}_K) := \sum_{k=1}^{K} \frac{1}{\tau} \mathcal{W}_{\text{deform}}[\phi_k, \mathcal{O}_{k-1}],$$

where $\phi_k : \overline{\mathcal{O}}_{k-1} \to \mathbb{R}^d$ is a minimizer of the deformation energy $\mathcal{W}_{\text{deform}}[\phi_k, \cdot]$ under the constraint $\phi_k(\mathcal{S}_{k-1}) = \mathcal{S}_k$. Furthermore, the discrete path length is defined as

$$\mathbf{L}_\tau(\mathcal{S}_0, \ldots, \mathcal{S}_K) := \sum_{k=1}^{K} \sqrt{\mathcal{W}_{\text{deform}}[\phi_k, \mathcal{O}_{k-1}]}\,.$$

Let us make a brief remark on the proper scaling factors. The deformation energy $\mathcal{W}_{\text{deform}}[\phi_k, \mathcal{O}_{k-1}]$ is expected to scale like τ^2 (cf. (7)). Hence, the factor $\frac{1}{\tau}$ ensures the discrete dissipation measure to be conceptually independent of the time step size. The same holds for the discrete length measure $\mathbf{L}_\tau(\mathcal{S}_0, \ldots, \mathcal{S}_K)$.

To ensure that the above-defined dissipation and length of discrete paths in shape space are well defined, a minimizing deformation ϕ_k of the elastic energy $\mathcal{W}_{\text{deform}}[\cdot, \mathcal{O}_{k-1}]$ with $\phi_k(\mathcal{S}_{k-1}) = \mathcal{S}_k$ has to exist. In fact, this holds for objects $\mathcal{O}_{k-1}$ and $\mathcal{O}_k$ with Lipschitz boundaries $\mathcal{S}_{k-1}$ and $\mathcal{S}_k$ for which there exists at least one bi-Lipschitz deformation $\hat{\phi}_k$ of $\mathcal{O}_{k-1}$ into $\mathcal{O}_k$ for $k = 1, \ldots, K$ [83].

With the notion of dissipation at hand, we can define a discrete geodesic path following the standard paradigms in differential geometry.

Definition 4 (Discrete geodesic path). A discrete path $\mathcal{S}_0, \mathcal{S}_1 \ldots, \mathcal{S}_K$ in a set of admissible shapes $\mathbf{S}$ connecting two shapes $\mathcal{S}_A = \mathcal{S}_0$ and $\mathcal{S}_B = \mathcal{S}_K$ is a discrete geodesic if there exists an associated family of deformations $(\phi_k)_{k=1,\ldots,K}$ such

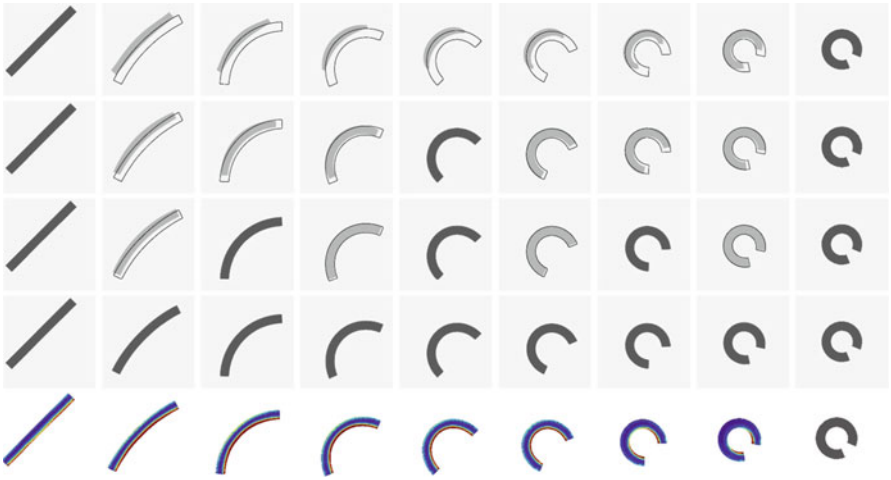


Fig. 19 Discrete geodesics between a straight and a rolled up bar, from first row to fourth row based on one, two, four, and eight time steps. The *light gray shapes* in the first, second, and third row show a linear interpolation of the deformations connecting the *dark gray shapes*. The shapes from the finest time discretization are overlayed over the others as *thin black lines*. In the last row, the rate of viscous dissipation is rendered on the shape domains $\mathcal{O}_1, \ldots, \mathcal{O}_7$ from the previous row, color coded as

that $(\phi_k, \mathcal{S}_k)_{k=1,\ldots,K}$ minimize the total energy $\sum_{k=1}^{K} \frac{1}{\tau} \mathcal{W}_{\text{deform}}[\tilde{\phi}_k, \tilde{\mathcal{O}}_{k-1}]$ over all intermediate shapes $\tilde{\mathcal{S}}_1 = \partial\tilde{\mathcal{O}}_1, \ldots, \tilde{\mathcal{S}}_{K-1} = \partial\tilde{\mathcal{O}}_{K-1} \in \mathbf{S}$ and all possible matching deformations $\tilde{\phi}_1, \ldots, \tilde{\phi}_K$ with $\tilde{\phi}_k(\tilde{\mathcal{S}}_{k-1}) = \tilde{\mathcal{S}}_k$ for $k = 1, \ldots, K$.

Examples of discrete geodesics are provided in Figs. 19 and 20. Apparently, the frame indifference and the (local) injectivity property of the matching deformations, which are ensured by the nonlinear deformation energy $\mathcal{W}_{\text{deform}}$, allow the computation of reasonable discrete geodesics with only few intermediate shapes. Under sufficient growth conditions on the integrand of the deformation energy $\mathcal{W}_{\text{deform}}$, the existence of discrete geodesics is guaranteed at least for certain compact sets $\mathbf{S}$ of admissible shapes, e.g., shapes $\mathcal{S}$ which can be described by spline curves with a finite set of control points from some compact domain and which satisfy a uniform cone condition in the sense that each $x \in \mathcal{S}$ is the tip of two cones with fixed height and opening angle which lie completely on either side of $\mathcal{S}$ [83]. Such requirements on $\mathbf{S}$ are necessary since the known regularity theory for deformation energies of the employed type does not allow to prove Lipschitz regularity of optimal deformations so that the intermediate shapes might degenerate.

The discrete dissipation as the sum of matching deformation energies indeed represents an approximation to the time-continuous dissipation of a velocity field from section "Path-Based, Viscous Riemannian Setup." If a smooth path in shape space is considered which is interpolated at discrete times $t_k = k\tau$, $k = 0, \ldots, K$

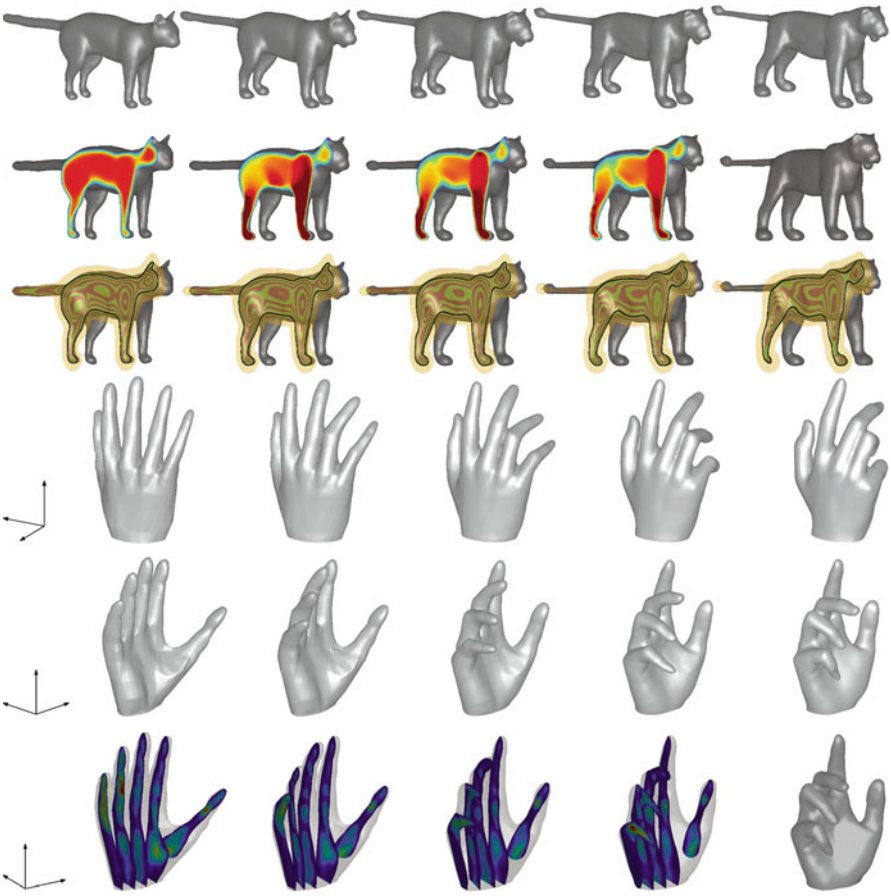


Fig. 20 Discrete geodesic between a cat and a lion and between the hand shapes m336 and m324 from the Princeton Shape Benchmark [72]. For both examples, the local dissipation is color coded on slices through the shapes as

and if for $t \in [t_{k-1}, t_k)$, $v_\tau(t) = \frac{(\phi_k - \mathbb{1})}{\tau} \circ \left(\frac{t_k - t}{\tau}\mathbb{1} + \frac{t - t_{k-1}}{\tau}\phi_k\right)^{-1}$ denotes the velocity field which generates the associated matching deformations ϕ_k, then as the time step size $\tau = \frac{1}{K}$ decreases and v_τ converges against a smooth velocity field v, the discrete dissipation converges against the time-continuous dissipation (2) induced by v (cf. [83] for details).

Within this framework of geodesics in shape space, the strict constraints that one shape is deformed exactly into another one are often inadequate in applications as has already been discussed in section "State-Based, Path-Independent Elastic Setup" for the state-based, elastic setup. For the computation of an elastic dissimilarity measure, the single matching constraint could be relaxed as a mismatch penalty. In the Riemannian, viscous setting, we pursue the same concept; however, the particular

Fig. 21 Discrete geodesic between the straight and the folded bar from Fig. 19, where the *black region* of the initial shape is constrained to be matched to the *black region* of the final shape. The *bottom row* shows a color coding of the corresponding viscous dissipation. Due to the strong change in relative position of the *black region*, the intermediate shapes exhibit a strong asymmetry and high dissipation near the bar ends

form of the employed constraints depends on the chosen view on shape geodesics. In the framework of geodesics as paths of diffeomorphisms, which we introduced at the beginning of this section, there is the single constraint $\phi(\mathcal{S}_A) = \mathcal{S}_B$, meaning that the induced diffeomorphism ϕ maps the initial shape $\mathcal{S}_A$ onto the final shape $\mathcal{S}_B$. This constraint can be relaxed in the same manner as in section "State-Based, Path-Independent Elastic Setup" via a penalty measuring the mismatch of the shapes or of the corresponding objects. For the time-discrete geodesic setting, we have a sequence of matching constraints $\phi_k(\mathcal{S}_{k-1}) = \mathcal{S}_k$, $k = 1, \ldots, K$, each of which can again be relaxed by the same means. In fact, we add to the discrete dissipation of a set $(\phi_k)_{k=1,\ldots,K}$ of deformations a sum of mismatch penalties $\sum_{k=1}^{K} \mathrm{vol}\left(\mathcal{O}_{k-1} \triangle \phi_k^{-1}(\mathcal{O}_k)\right)$. In the limit for vanishing time step size $\tau = \frac{1}{K}$ and under the same conditions as above, this sum can be shown to converge against the optical flow-type functional $\int_{\mathcal{T}} |(1, v(t)) \cdot n\,[t, \mathcal{S}(t)]|\,\mathrm{d}a$ for the unit outward normal $n[t, \mathcal{S}(t)]$ to the space time shape tube $\mathcal{T} = \bigcup_{t\in[0,1]} \{t\} \times \mathcal{S}(t)$. Furthermore, $\sum_{k=1}^{K} \tau \mathcal{L}[\mathcal{S}_k]$ with $\mathcal{L}[\mathcal{S}_k] = \mathcal{H}^{d-1}(\mathcal{S}_k)$ has been employed as regularization, which in the limit for $\tau \to 0$ converges against the integral $\int_0^1 \mathcal{H}^{d-1}(\mathcal{S}(t))\,\mathrm{d}t$.

Real-world objects are most often not only characterized by their outer contour but also contain internal structures that have to be matched properly when computing the similarity between two objects. As an example, consider the straight and the folded rod in Fig. 21. The rods consist of three distinct components, which imposes a constraint on reasonable connecting paths: Each component is to be mapped onto its correct counterpart. A shortest path under this constraint obviously differs significantly from the geodesic which just matches the outer contours (cf. Fig. 19).

This observation calls for a generalization of shapes, an example of which we have already seen in the context of an elastic shape space in Fig. 18, where the edge set of an image was considered as a shape. Here, let us adopt a slightly different approach and regard shapes as being composed of a number of subcomponents. In detail, instead of a geodesic between just two shapes $\mathcal{S}_A = \partial\mathcal{O}_A$ and $\mathcal{S}_B = \partial\mathcal{O}_B$, we now seek a geodesic path $\left(\mathcal{S}^i(t)\right)_{i=1,\ldots,m}$ with $\mathcal{S}^i(t) = \partial\mathcal{O}^i(t)$ for $t \in [0, 1]$, between two collections of m separate shapes, $\left(\mathcal{S}_A^i\right)_{i=1,\ldots,m}$ with $\mathcal{S}_A^i(t) = \partial\mathcal{O}_A^i(t)$ and $\left(\mathcal{S}_B^i\right)_{i=1,\ldots,m}$ with $\mathcal{S}_B^i(t) = \partial\mathcal{O}_B^i(t)$. The geodesic path is supposed to

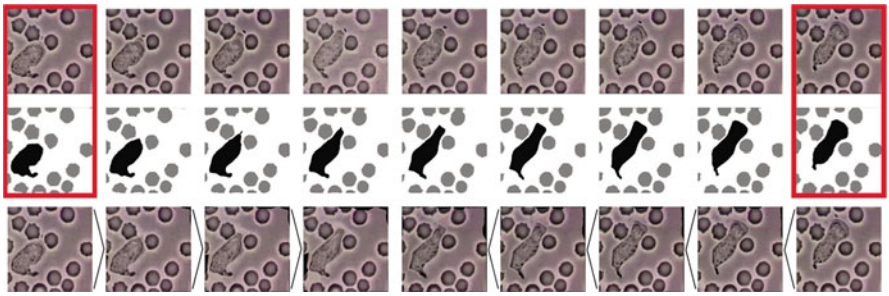


Fig. 22 *Top*: frames from a real video sequence of a white blood cell among a number of *red* ones (Courtesy of Robert A. Freitas, Institute for Molecular Manufacturing, California, USA). *Middle*: computed discrete geodesic between the segmented shapes in the first and the last frame. *Bottom*: pushforward of the initial (first four shapes) and pullback of the final frame (last five shapes) according to the geodesic flow

be generated by a joint motion field $v(t) : \bigcup_{i=1}^{m} \mathcal{O}^i(t) \to \mathbb{R}^d$. The single objects $\mathcal{O}^i(t)$ can then be regarded as the subcomponents of an overall object $\bigcup_{i=1}^{m} \mathcal{O}^i(t)$. The total dissipation along the path is measured exactly as before by

$$\mathbf{Diss}\left[\left(v(t), (\mathcal{O}^i(t))_{i=1,\dots,m}\right)_{t\in[0,1]}\right] = \int_0^1 \int_{\bigcup_{i=1}^{m} \mathcal{O}^i(t)} \frac{\lambda}{2} \left(\mathrm{tr}\epsilon[v]\right)^2 + \mu \,\mathrm{tr}\left(\epsilon[v]^2\right) \mathrm{d}x \,\mathrm{d}t \,.$$

This naturally translates to the discrete dissipation of a path with $K+1$ intermediate shape collections $\left(\mathcal{S}_k^i\right)_{i=1,\dots,m}$, $k = 0, \dots, K$,

$$\sum_{k=1}^{K} \mathcal{W}_{\mathrm{deform}}\left[\phi_k, \left(\mathcal{O}_{k-1}^i\right)_{i=1,\dots,m}\right] := \sum_{k=1}^{K} \int_{\bigcup_{i=1}^{m} \mathcal{O}_{k-1}^i} W(\mathcal{D}\phi_k)\,\mathrm{d}x,$$

where the deformations ϕ_k satisfy the constraints $\phi_k\left(\mathcal{S}_{k-1}^i\right) = \mathcal{S}_k^i$ for $k = 1, \dots, K$, $i = 1, \dots, m$, and $\mathcal{S}_0^i = \mathcal{S}_A^i$, $\mathcal{S}_K^i = \mathcal{S}_B^i$, $i = 1, \dots, m$.

The different object components can of course be assigned different material properties. Figure 22 shows frames from a real video sequence of moving white and red blood cells (top) as well as a discrete geodesic between the first and last frame (middle) for which the material parameters of the white blood cell were chosen twenty times weaker than for the red blood cells. The result is a nonlinear interpolation between distant frames which is in good agreement with the actually observed motion. Once geodesic distances between shapes are defined, one can statistically analyze ensembles of shapes and cluster them in groups based on the geodesic distance as a reliable measure for the similarity of shapes. Two exemplary examples are provided by the evaluation of geodesic distances between different 2D letters (Fig. 23, *left*) and between six different 3D foot shapes (Fig. 23, *right*). In the 2D example, we clearly identify three distinct clusters (*B*s, *X*s, and *M*).

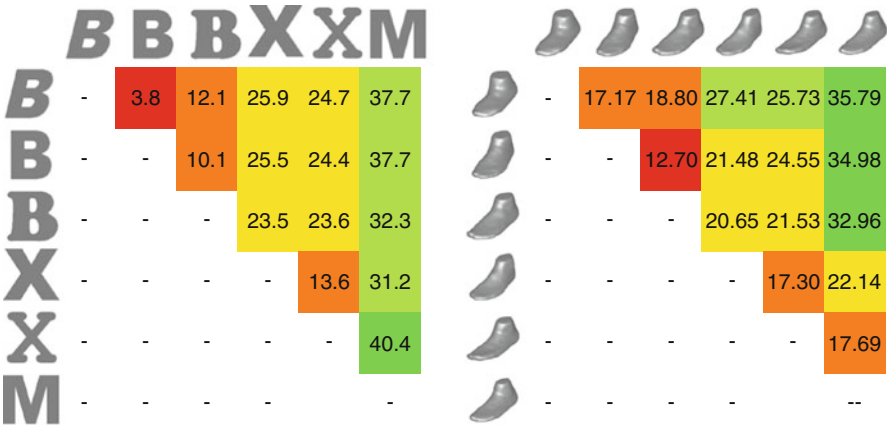


Fig. 23 *Left*: pairwise geodesic distances between (also topologically) different letter shapes. *Right*: pairwise geodesic distances between different scanned 3D feet. The feet have volumes 499.5, 500.6, 497.6, 434.7, 432, and 381 cm^3, respectively

A Collection of Computational Tools

So far, we have investigated some of the many aspects on mathematical models in shape space without any discussion of the corresponding computational tools and numerical algorithms. Hence, let us at least briefly mention some fundamental computational aspects to effectively deal with general classes of shapes as boundary contours of volumetric objects.

At first, we replace the strict separation between material inside the object and void outside by substituting the void with a material which is several orders of magnitude softer than inside the object. This relaxation is important with respect to the existence analysis and the stabilization of the computational method. In fact, we replace the deformation energy $\mathcal{W}_{\text{deform}}[\phi, \mathcal{O}] = \int_{\mathcal{O}} W(\mathcal{D}\phi)\,\mathrm{d}x$ by the energy $\mathcal{W}^{\eta}_{\text{deform}}[\phi, \mathcal{O}] = \int_{\Omega} \big((1-\eta)\chi_{\mathcal{O}} + \eta\big)\, W(\mathcal{D}\phi)\,\mathrm{d}x$ for a small constant η. In the implementation which underlies the above applications, for $\eta = 10^{-4}$ one observes no significant qualitative impact of this regularization on the solution. Furthermore, as mentioned above, to ensure regularity of the shape contour $\mathcal{S}$, we take into account the area functional $\mathcal{L}[\mathcal{S}] = \int_{\mathcal{S}} \mathrm{d}a$ as a prior, weighted with a small factor.

Compared to a parametric description of shapes, e.g., as a polygonal line or a triangulated surface, an implicit description has several advantages. In particular, it does not require a remeshing even in case of large deformations, it allows for topological transitions without any extra handling of the associated singularities, and it can be combined with multi-scale relaxation schemes for an efficient minimization of the involved functionals.

In what follows, we consider a level set and a phase field description of shapes and outline the general framework of a multi-scale method based on finite element calculus. In fact, the phase field model has been used in the examples for the

elastic shape averaging and the PCA, whereas the level set method has served as a numerical building block for the computation of time-discrete shape geodesics.

Shapes Described by Level Set Functions

The level set method first presented by Osher and Sethian [60] has been used for a wide range of applications [59, 71]. Burger and Osher gave an overview in the context of shape optimization [7]. To numerically solve variational problems in shape space, we assume a shape $\mathcal{S}$ to be represented by the zero level set $\{x \in \Omega : u(x) = 0\}$ of a scalar function $u : \Omega \to \mathbb{R}$ on a computational domain $\Omega \subset \mathbb{R}^d$. Furthermore, the zero super level set $\{x \in \Omega : u(x) > 0\}$ defines the corresponding object domain $\mathcal{O}$. This shape description can be incorporated in a variational approach following the approximation proposed by Chan and Vese [9]. In fact, the partition of the domain Ω into object and background is encoded via a regularized Heaviside function $H_\varepsilon \circ u$. As in [9], we consider the function $H_\varepsilon(x) := \frac{1}{2} + \frac{1}{\pi} \arctan\left(\frac{x}{\varepsilon}\right)$, where ε is a scale parameter representing the width of the smeared-out shape contour. Then, a deformation energy $\mathcal{W}^{\eta}_{\text{deform}}[\phi, \mathcal{O}] = \int_\Omega \left((1-\eta)\chi_{\mathcal{O}} + \eta\right) W(\mathcal{D}\phi)\, \mathrm{d}x$ is approximated by

$$\mathcal{W}^{\varepsilon,\eta}_{\text{deform}}[\phi, u] = \int_\Omega \left((1-\eta) H_\varepsilon(u) + \eta\right) W(\mathcal{D}\phi)\, \mathrm{d}x.$$

Furthermore, the energy $\mathcal{F}[\mathcal{S}_A, \phi, \mathcal{S}_B] = \mathrm{vol}\left(\mathcal{O}_A \triangle \phi^{-1}(\mathcal{O}_B)\right)$ measuring the volumetric mismatch between an object $\mathcal{O}_A$ and the pullback of an object $\mathcal{O}_B$ under a deformation ϕ can be approximated by

$$\mathcal{F}^{\varepsilon}[u_A, \phi, u_B] = \int_\Omega \left(H_\varepsilon(u_B \circ \phi) - H_\varepsilon(u_A)\right)^2\, \mathrm{d}x,$$

where u_A, u_B are level set representations of the shapes $\mathcal{S}_A$ and $\mathcal{S}_B$, respectively. Finally, the surface area of a shape $\mathcal{S}$, which appears as a prior, is replaced by the total variation of $H_\varepsilon \circ u$, and we obtain

$$\mathcal{L}^{\varepsilon}[u] = \int_\Omega |\nabla H_\varepsilon(u)|\, \mathrm{d}x.$$

Let us emphasize that in the actual energy minimization algorithm, the guidance of an initial zero level set toward the final shape relies on the nonlocal support of the derivative of the regularized Heaviside function (cf. [8]).

Shapes Described via Phase Fields

An alternative to a level set description of shapes is a phase field representation. Physically, the phase field approach is inspired by the observation that interfaces are usually not sharp but characterized by a diffusive transition. Mathematically, there are two basic types of such phase field representations, a single phase approach as the one presented by Ambrosio and Tortorelli [1] for the approximation of the

Mumford–Shah model [56] and the double phase approach by Modica and Mortola [55] used to approximate surface integrals. In the shape context studied here, let us focus on the single phase model. Thus, a shape $\mathcal{S}$ is encoded by a continuous, piecewise smooth phase field function $u : \Omega \to \mathbb{R}$ which is zero on $\mathcal{S}$, but close to one everywhere else. The specific profile of the phase field function u for a shape $\mathcal{S}$ is determined via the phase field approximation

$$\mathcal{L}^\varepsilon[u] = \frac{1}{2}\int_\Omega \varepsilon|\nabla u|^2 + \frac{1}{\varepsilon}(u-1)^2 \,\mathrm{d}x$$

of the involved surface area $\int_{\mathcal{S}} \mathrm{d}a$. As in the above level set model, the phase field parameter ε determines the width of the diffusive interface. Different from the level set model by Chan and Vese, the interface profile is not explicitly prescribed but implicitly encoded in the variational approach as the profile attained by minimizers of the functional. Based on this phase field model the penalty functional $\mathcal{F}[\mathcal{S}_A, \phi, \mathcal{S}_B] = \mathcal{H}^{d-1}(\mathcal{S}_A \triangle \phi^{-1}(\mathcal{S}_B))$ measuring the area mismatch between a shape $\mathcal{S}_A$ and the pullback of a shape $\mathcal{S}_B$ under a deformation ϕ can be approximated by

$$\mathcal{F}^\varepsilon[u_A, \phi, u_B] = \frac{1}{\varepsilon}\int_\Omega (u_B \circ \phi)^2(1-u_A)^2 + u_A^2(1-u_B\circ\phi)^2 \,\mathrm{d}x,$$

where u_A, u_B are phase fields representing the shapes $\mathcal{S}_A$ and $\mathcal{S}_B$, respectively. In this type of models, the deformation energy $\mathcal{W}^\eta_{\text{deform}}[\phi, \mathcal{O}]$ cannot be realized based on a phase field function u due to the fact that a single phase model allows to identify the shape itself but does not distinguish its inside and outside. Therefore, in the presented applications of elastic shape averaging and the elastic PCA, the input objects and thus their characteristic functions $\chi_{\mathcal{O}}$ were given a priori.

Multi-scale Finite Element Approximation

For the spatial discretization of the functionals in the above variational approaches, the finite element method can be applied. Hence, the level set function or the phase field u, representing a (unknown) shape $\mathcal{S}$, and the different components of the deformations ϕ are represented by continuous, piecewise multilinear (trilinear in 3D and bilinear in 2D) finite element functions U and Φ on a regular grid superimposed on the domain $\Omega = [0,1]^d$. For the ease of implementation, a dyadic grid resolution with $2^L + 1$ vertices in each direction and a grid size $h = 2^{-L}$ is chosen.

Descent algorithm. The functionals depend nonlinearly both on the discrete deformations Φ (due to the concatenation $U \circ \Phi$ and the nonlinear integrand $W(\cdot)$ of the deformation energy) as well as on the discrete level set or phase field functions U (e.g., due to the concatenation of the level set function with the regularized Heaviside function $H_\varepsilon(\cdot)$). In our energy relaxation algorithm for fixed grid size, we employ a gradient descent approach. We constantly alternate between performing a

single gradient descent step for all deformations and the level set or phase field functions.

Numerical quadrature. Integral evaluations in the descent algorithm are performed by Gaussian quadrature of third order on each grid cell. For various terms, we have to evaluate pullbacks $U \circ \Phi$ of a discretized level set function U or a test function under a discretized deformation Φ. Let us emphasize that quadrature based on nodal interpolation of $U \circ \Phi$ would lead to artificial displacements near the shape edges accompanied by strong artificial tension. Hence, in our algorithm, if $\Phi(x)$ lies inside Ω for a quadrature point x, then the pullback is evaluated exactly at x. Otherwise, we project $\Phi(x)$ back onto the boundary of Ω and evaluate U at that projection point.

Cascadic multi-scale algorithm. The variational problem considered here is highly nonlinear, and for fixed time step size, the proposed scheme is expected to have very slow convergence; also it might end up in some nearby local minimum. Here, a multilevel approach (initial optimization on a coarse scale and successive refinement) turns out to be indispensable in order to accelerate convergence and not to be trapped in undesirable local minima. Due to our assumption of a dyadic resolution 2^L+1 in each grid direction, we are able to build a hierarchy of grids with $2^l + 1$ nodes in each direction for $l = L, \ldots, 0$. Via a simple restriction operation, we project every finite element function to any of these coarse grid spaces. Starting the optimization on a coarse grid, the results from coarse scales are successively prolongated onto the next grid level for a refinement of the solution [5]. Hence, the construction of a grid hierarchy allows to solve coarse scale problems in our multi-scale approach on coarse grids. Since the width ε of the diffusive shape representation should naturally scale with the grid width h, we choose $\varepsilon = h$.

5 Conclusion

Let us close with a comparison of path- and state-based shape space. Already in section "Conceptual Differences Between the Path- and State-Based Dissimilarity Measures," we have studied the difference between the state-based dissimilarity measure d_{elast} and the path-based distance d_{viscous}. Based on the applications considered in the previous sections, let us compare the underlying concepts now more on a conceptual level of the geometry of shape space:

- *Non-uniqueness of shape averages*. Due to the nonlinearity of the elastic variational problem, local minimizers of the elastic energy might be nonunique. There might even exist different minimizing deformations with the same elastic energy. Mechanically, this nonuniqueness is frequently associated with different buckling modes, which occur in case of large, geometrically nonlinear deformations. Hence, the shape average need not be uniquely defined, except in the small displacement case, where a linear elastic model (8) applies. In case of the path-

based approach, geodesics (shortest) do not have to be unique either. Indeed, a geodesic is the unique shortest path only until the first conjugate point. Hence, the shape average is in a strict sense not well defined if the distances are sufficiently large.

- *Different physical interpretation of the PCA*. In the Riemannian setup with the metric being the rate of viscous dissipation, the $\log_{\mathcal{S}} \mathcal{S}_k$ corresponds to the initial velocity $v_k : \mathcal{S} \to \mathbb{R}^d$ in the (optimal transport) flow of $\mathcal{O}$ associated with shape $\mathcal{S}$ into $\mathcal{O}_k$ associated with the kth input shape $\mathcal{S}_k$. In the elastic model, the boundary stress $\sigma_k \nu : \partial\mathcal{O} \to \mathbb{R}^d$ results from the deformation ϕ_k of $\mathcal{O}_k$ onto the average object $\mathcal{O}$ and effectively is the restoring force acting on the average shape $\mathcal{S}$. Via the linearized elasticity problem in the prestressed compound configuration of the average object $\mathcal{O}$, these restoring forces are identified with displacements u_k. Depending on the model, either the flow velocities v_k or the linear elastic displacements u_k form the basis of a covariance analysis in the linear vector space of mappings $\overline{\mathcal{O}} \to \mathbb{R}^d$. The outcome are principal shape variations of the average shape, either generated by motion fields or displacements, respectively.
- *Quantitative shape analysis*. The Riemannian metric given by the rate of viscous dissipation in the path-based viscous fluid approach allows direct comparison of multiple ensembles of shapes via pairwise distance computations. Due to the lack of a triangle inequality, this is possible only in a restricted sense in the state-based elastic approach, where dissimilarity measures for one fixed shape and a set of varying shapes can be computed.
- The method of choice depends on the *specific application*. If shapes are considered as boundaries of objects with a viscous fluid inside, then the path-based approach would be more appropriate. The state-based elastic approach is favorable for objects which behave more like deformable solids.

Acknowledgments The model proposed in section "Viscous Fluid-Based Shape Space" has been developed in cooperation with Leah Bar and Guillermo Sapiro from the University of Minnesota. Benedikt Wirth has been funded by the Bonn International Graduate School in Mathematics. Furthermore, the work was supported by the Deutsche Forschungsgemeinschaft, SPP 1253 "Optimization with Partial Differential Equations." Part of Figs. 3–4 and 19–23 have been taken from [83], the results from Figs. 6, 8, and 10–18 stem from [67, 69].

Cross-References

▸ Level Set Methods for Structural Inversion and Image Reconstruction
▸ Mumford and Shah Model and Its Applications to Image Segmentation and Image Restoration
▸ Numerical Methods and Applications in Total Variation Image Restoration
▸ Shape Spaces

References

1. Ambrosio, L., Tortorelli, V.M.: On the approximation of free discontinuity problems. B. Unione Mat. Ital. B **6**(7), 105–123 (1992)
2. Ball, J.: Global invertibility of Sobolev functions and the interpenetration of matter. Proc. R. Soc. Edinb. **88A**, 315–328 (1981)
3. Beg, M.F., Miller, M.I., Trouvé, A., Younes, L.: Computing large deformation metric mappings via geodesic flows of diffeomorphisms. Int. J. Comput. Vis. **61**(2), 139–157 (2005)
4. Berkels, B., Linkmann, G., Rumpf, M.: An $SL(2)$ invariant shape median (2009, submitted)
5. Bornemann, F., Deuflhard, P.: The cascadic multigrid method for elliptic problems. Numer. Math. **75**(2), 135–152 (1996)
6. Bronstein, A., Bronstein, M., Kimmel, R.: Numerical Geometry of Non-rigid Shapes. Monographs in Computer Science. Springer, New York (2008)
7. Burger, M., Osher, S.J.: A survey on level set methods for inverse problems and optimal design. Eur. J. Appl. Math. **16**(2), 263–301 (2005)
8. Caselles, V., Kimmel, R., Sapiro, G.: Geodesic active contours. Int. J. Comput. Vis. **22**(1), 61–79 (1997)
9. Chan, T.F., Vese, L.A.: Active contours without edges. IEEE Trans. Image Process. **10**(2), 266–277 (2001)
10. Charpiat, G., Faugeras, O., Keriven, R.: Approximations of shape metrics and application to shape warping and empirical shape statistics. Found. Comput. Math. **5**(1), 1–58 (2005)
11. Charpiat, G., Faugeras, O., Keriven, R., Maurel, P.: Distance-based shape statistics. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2006), Toulouse, vol. 5, pp. 925–928 (2006)
12. Chen, S.E., Parent, R.E.: Shape averaging and its applications to industrial design. IEEE Comput. Graph. Appl. **9**(1), 47–54 (1989)
13. Chorin, A.J., Marsden, J.E.: A Mathematical Introduction to Fluid Mechanics. Volume 4 of Texts in Applied Mathematics. Springer, New York (1990)
14. Christensen, G.E., Rabbitt, R.D., Miller, M.I.: 3D brain mapping using a deformable neuroanatomy. Phys. Med. Biol. **39**(3), 609–618 (1994)
15. Ciarlet, P.G.: Three-Dimensional Elasticity. Elsevier Science B.V., New York (1988)
16. Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J.: Active shape models–their training and application. Comput. Vis. Image Underst. **61**(1), 38–59 (1995)
17. Cremers, D., Kohlberger, T., Schnörr, C.: Shape statistics in kernel space for variational image segmentation. Pattern Recognit. **36**, 1929–1943 (2003)
18. Dacorogna, B.: Direct Methods in the Calculus of Variations. Springer, New York (1989)
19. Dambreville, S., Rathi, Y., Tannenbaum, A.: A shape-based approach to robust image segmentation. In: Campilho, A., Kamel, M. (eds.) IEEE Computer Society Conference on Computer Vision and Pattern Recognition, New York. Volume 4141 of LNCS, pp. 173–183 (2006)
20. Delfour, M.C., Zolésio, J.: Geometries and Shapes: Analysis, Differential Calculus and Optimization. Advance in Design and Control, vol. 4. SIAM, Philadelphia (2001)
21. do Carmo, M.P.: Riemannian Geometry. Birkhäuser, Boston (1992)
22. Droske, M., Rumpf, M.: Multi scale joint segmentation and registration of image morphology. IEEE Trans. Pattern Recognit. Mach. Intell. **29**(12), 2181–2194 (2007)
23. Dupuis, D., Grenander, U., Miller, M.: Variational problems on flows of diffeomorphisms for image matching. Q. Appl. Math. **56**, 587–600 (1998)
24. Eckstein, I., Pons, J.P., Tong, Y., Kuo, C.C., Desbrun, M.: Generalized surface flows for mesh processing. In: Eurographics Symposium on Geometry Processing, Barcelona (2007)
25. Elad (Elbaz), A., Kimmel, R.: On bending invariant signatures for surfaces. IEEE Trans. Pattern Anal. Mach. Intell. **25**(10), 1285–1295 (2003)

26. Fletcher, P.T., Lu, C., Joshi, S.: Statistics of shape via principal geodesic analysis on Lie groups. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Los Alamitos, vol. 1. pp. 95–101 (2003)
27. Fletcher, P., Lu, C., Pizer, S., Joshi, S.: Principal geodesic analysis for the study of nonlinear statistics of shape. IEEE Trans. Med. Imaging **23**(8), 995–1005 (2004)
28. Fletcher, T., Venkatasubramanian, S., Joshi, S.: Robust statistics on Riemannian manifolds via the geometric median. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Anchorage (2008)
29. Fletcher, P., Whitaker, R.: Riemannian metrics on the space of solid shapes. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2006, Copenhagen (2006)
30. Fréchet, M.: Les éléments aléatoires de nature quelconque dans un espace distancié. Ann. Inst. H. Poincaré **10**, 215–310 (1948)
31. Fuchs, M., Jüttler, B., Scherzer, O., Yang, H.: Shape metrics based on elastic deformations. J. Math. Imaging Vis. **35**(1), 86–102 (2009)
32. Fuchs, M., Scherzer, O.: Segmentation of biologic image data with a-priori knowledge. FSP report, Forschungsschwerpunkt S92 52, Universität Innsbruck, Innsbruck (2007)
33. Fuchs, M., Scherzer, O.: Regularized reconstruction of shapes with statistical a priori knowledge. Int. J. Comput. Vis. **79**(2), 119–135 (2008)
34. Glaunès, J., Qiu, A., Miller, M.I., Younes, L.: Large deformation diffeomorphic metric curve mapping. Int. J. Comput. Vis. **80**(3), 317–336 (2008)
35. Hafner, B., Zachariah, S., Sanders, J.: Characterisation of three-dimensional anatomic shapes using principal components: application to the proximal tibia. Med. Biol. Eng. Comput. **38**, 9–16 (2000)
36. Hong, B.W., Soatto, S., Vese, L.: Enforcing local context into shape statistics. Adv. Comput. Math. (online first) (2008)
37. Joshi, S., Davis, B., Jomier, M., Gerig, G.: Unbiased diffeomorphic atlas construction for computational anatomy. NeuroImage **23**(Suppl. 1), 151–160 (2004)
38. Karcher, H.: Riemannian center of mass and mollifier smoothing. Commun. Pure Appl. Math. **30**(5), 509–541 (1977)
39. Kendall, D.G.: Shape manifolds, procrustean metrics, and complex projective spaces. Bull. Lond. Math. Soc. **16**, 81–121 (1984)
40. Kilian, M., Mitra, N.J., Pottmann, H.: Geometric modeling in shape space. ACM Trans. Graph. **26**(64), 1–8 (2007)
41. Klassen, E., Srivastava, A., Mio, W., Joshi, S.H.: Analysis of planar shapes using geodesic paths on shape spaces. IEEE Trans. Pattern Anal. Mach. Intell. **26**(3), 372–383 (2004)
42. Klingenberg, W.P.A.: Riemannian Geometry. Walter de Gruyter, Berlin (1995)
43. Leventon, M.E., Grimson, W.E.L., Faugeras, O.: Statistical shape influence in geodesic active contours. In: 5th IEEE EMBS International Summer School on Biomedical Imaging, Berder Island (2002)
44. Ling, H., Jacobs, D.W.: Shape classification using the inner-distance. IEEE Trans. Pattern Anal. Mach. Intell. **29**(2), 286–299 (2007)
45. Liu, X., Shi, Y., Dinov, I., Mio, W.: A computational model of multidimensional shape. Int. J. Comput. Vis. (online first) (2010)
46. Manay, S., Cremers, D., Hong, B.W., Yezzi, A.J., Soatto, S.: Integral invariants for shape matching. IEEE Trans. Pattern Anal. Mach. Intell. **28**(10), 1602–1618 (2006)
47. Marsden, J.E., Hughes, T.J.R.: Mathematical Foundations of Elasticity. Prentice-Hall, Englewood Cliffs (1983)
48. McNeill, G., Vijayakumar, S.: 2d shape classification and retrieval. In: Proceedings of the 19th International Joint Conference on Artificial Intelligence, Edinburgh, pp. 1483–1488 (2005)
49. Mémoli, F.: Gromov-Hausdorff distances in Euclidean spaces. In: Workshop on Non-rigid Shape Analysis and Deformable Image Alignment (CVPR Workshop, NORDIA'08), Anchorage (2008)
50. Mémoli, F., Sapiro, G.: A theoretical and computational framework for isometry invariant recognition of point cloud data. Found. Comput. Math. **5**, 313–347 (2005)

51. Michor, P.W., Mumford, D.: Riemannian geometries on spaces of plane curves. J. Eur. Math. Soc. **8**, 1–48 (2006)
52. Michor, P.W., Mumford, D., Shah, J., Younes, L.: A metric on shape space with explicit geodesics. Rend. Lincei Mat. Appl. **9**, 25–57 (2008)
53. Miller, M., Trouvé, A., Younes, L.: On the metrics and Euler-Lagrange equations of computational anatomy. Annu. Rev. Biomed. Eng. **4**, 375–405 (2002)
54. Miller, M.I., Younes, L.: Group actions, homeomorphisms, and matching: a general framework. Int. J. Comput. Vis. **41**(1–2), 61–84 (2001)
55. Modica, L., Mortola, S.: Un esempio di-Γ^--convergenza. Boll. Un. Mat. Ital. B (5) **14**(1), 285–299 (1977)
56. Mumford, D., Shah, J.: Optimal approximation by piecewise smooth functions and associated variational problems. Commun. Pure Appl. Math. **2**, 577–685 (1989)
57. Nečas, J., Čsilhavý, M.: Multipolar viscous fluids. Q. Appl. Math. **49**(2), 247–265 (1991)
58. Ogden, R.W.: Non-linear Elastic Deformations. Wiley, New York (1984)
59. Osher, S., Fedkiw, R.: Level Set Methods and Dynamic Implicit Surfaces. Volume 153 of Applied Mathematical Sciences. Springer, New York (2003)
60. Osher, S., Sethian, J.A.: Fronts propagating with curvature dependent speed: algorithms based on Hamilton–Jacobi formulations. J. Comput. Phys. **79**(1), 12–49 (1988)
61. Pennec, X.: Left-invariant Riemannian elasticity: a distance on shape diffeomorphisms? In: Mathematical Foundations of Computational Anatomy – MFCA 2006, Copenhagen, pp. 1–14 (2006)
62. Pennec, X., Stefanescu, R., Arsigny, V., Fillard, P., Ayache, N.: Riemannian elasticity: a statistical regularization framework for non-linear registration. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2005, Palm Springs. LNCS, pp. 943–950 (2005)
63. Perperidis, D., Mohiaddin, R., Rueckert, D.: Construction of a 4d statistical atlas of the cardiac anatomy and its use in classification. In: Duncan, J., Gerig, G. (eds.) Medical Image Computing and Computer Assisted Intervention, Palm Springs. Volume 3750 of LNCS, pp. 402–410 (2005)
64. Rathi, Y., Dambreville, S., Tannenbaum, A.: Statistical shape analysis using kernel PCA. Proc. SPIE **6064**, 425–432 (2006)
65. Rathi, Y., Dambreville, S., Tannenbaum, A.: Comparative analysis of kernel methods for statistical shape learning. In: Beichel, R., Sonka, M. (eds.) Computer Vision Approaches to Medical Image Analysis, Graz. Volume 4241 of LNCS, pp. 96–107 (2006)
66. Rueckert, D., Frangi, A.F., Schnabel, J.A.: Automatic construction of 3-d statistical deformation models of the brain using nonrigid registration. IEEE Trans. Med. Imaging **22**(8), 1014–1025 (2003)
67. Rumpf, M., Wirth, B.: A nonlinear elastic shape averaging approach. SIAM J. Imaging Sci. **2**(3), 800–833 (2009)
68. Rumpf, M., Wirth, B.: An elasticity approach to principal modes of shape variation. In: Proceedings of the Second International Conference on Scale Space Methods and Variational Methods in Computer Vision (SSVM 2009), Voss. Volume 5567 of LNCS, pp. 709–720 (2009)
69. Rumpf, M., Wirth, B.: An elasticity-based covariance analysis of shapes. Int. J. Comput. Vis. (2009, accepted)
70. Schmidt, F.R., Clausen, M., Cremers, D.: Shape matching by variational computation of geodesics on a manifold. In: Pattern Recognition, Berlin. Volume 4174 of LNCS, pp. 142–151. Springer, Berlin (2006)
71. Sethian, J.A.: Level Set Methods and Fast Marching Methods. Cambridge University Press, Cambridge (1999)
72. Shilane, P., Min, P., Kazhdan, M., Funkhouser, T.: The Princeton shape benchmark. In: Proceedings of the Shape Modeling International, Genova, pp. 167–178 (2004)
73. Söhn, M., Birkner, M., Yan, D., Alber, M.: Modelling individual geometric variation based on dominant eigenmodes of organ deformation: implementation and evaluation. Phys. Med. Biol. **50**, 5893–5908 (2005)

74. Spivak, M.: A Comprehensive Introduction to Differential Geometry, vol. 1. Publish or Perish, Boston (1970)
75. Srivastava, A., Jain, A., Joshi, S., Kaziska, D.: Statistical shape models using elastic-string representations. In: Narayanan, P. (ed.) Asian Conference on Computer Vision, Hyderabad. Volume 3851 of LNCS, pp. 612–621. Springer, Heidelberg (2006)
76. Sundaramoorthi, G., Yezzi, A., Mennucci, A.: Sobolev active contours. Int. J. Comput. Vis. **73**(3), 345–366 (2007)
77. Thorstensen, N., Segonne, F., Keriven, R.: Pre-image as karcher mean using diffusion maps: application to shape and image denoising. In: Proceedings of the Second International Conference on Scale Space Methods and Variational Methods in Computer Vision (SSVM 2009), Voss. Volume 5567 of LNCS, pp. 721–732 (2009)
78. Truesdell, C., Noll, W.: The Non-linear Field Theories of Mechanics. Springer, Berlin (2004)
79. Tsai, A., Yezzi, A., Wells, W., Tempany, C., Tucker, D., Fan, A., Grimson, W.E., Willsky, A.: A shape-based approach to the segmentation of medical imagery using level sets. IEEE Trans. Med. Imaging **22**(2), 137–154 (2003)
80. Vaillant, M., Glaunès, J.: Surface matching via currents. In: IPMI 2005: Information Processing in Medical Imaging, Glenwood Springs. Volume 3565 of LNCS, pp. 381–392. Springer (2005)
81. Wirth, B.: Variational methods in shape space. Dissertation, University Bonn, Bonn (2009)
82. Wirth, B., Bar, L., Rumpf, M., Sapiro, G.: Geodesics in shape space via variational time discretization. In: Proceedings of the 7th International Conference on Energy Minimization Methods in Computer Vision and Pattern Recognition (EMMCVPR'09), Bonn. Volume 5681 of LNCS, pp. 288–302 (2009)
83. Wirth, B., Bar, L., Rumpf, M., Sapiro, G.: A continuum mechanical approach to geodesics in shape space (2010, submitted to IJCV)
84. Yezzi, A.J., Mennucci, A.: Conformal metrics and true "gradient flows" for curves. In: ICCV 2005: Proceedings of the 10th IEEE International Conference on Computer Vision, Beijing, pp. 913–919 (2005)
85. Younes, L.: Computable elastic distances between shapes. SIAM J. Appl. Math. **58**(2), 565–586 (1998)
86. Younes, L., Qiu, A., Winslow, R.L., Miller, M.I.: Transport of relational structures in groups of diffeomorphisms. J. Math. Imaging Vis. **32**(1), 41–56 (2008)
87. Yushkevich, P., Fletcher, P.T., Joshi, S., Thalla, A., Pizer, S.M.: Continuous medial representations for geometric object modeling in 2d and 3d. Image Vis. Comput. **21**(1), 17–27 (2003)
88. Zolésio, J.P.: Shape topology by tube geodesic. In: IFIP Conference on System Modeling and Optimization, No. 21, pp. 185–204 (2004)

Manifold Intrinsic Similarity

Alexander M. Bronstein and Michael M. Bronstein

Contents

1 Introduction 1860
Problems 1861
Methods 1862
Chapter Outline 1862
2 Shapes as Metric Spaces 1862
Basic Notions 1863
Euclidean Geometry 1864
Riemannian Geometry 1864
Diffusion Geometry 1866
Diffusion Distances 1868
3 Shape Discretization 1869
Sampling 1869
Shape Representation 1872
4 Metric Discretization 1873
Shortest Paths on Graphs 1873
Fast Marching 1875
Diffusion Distance 1880
5 Invariant Shape Similarity 1882
Rigid Similarity 1883
Canonical Forms 1885
Gromov–Hausdorff Distance 1888
Graph-Based Methods 1891
Gromov–Wasserstein Distances 1892
Shape DNA 1893
6 Partial Similarity 1893
Significance 1893
Regularity 1894
Partial Similarity Criterion 1895

A.M. Bronstein (✉) • M.M. Bronstein
Computer Science Department, Technion-Israel Institute of Technology, Haifa, Israel
e-mail: bron@cs.technion.ac.il

© Springer Science+Business Media New York 2015

O. Scherzer (ed.), *Handbook of Mathematical Methods in Imaging*,
DOI 10.1007/978-1-4939-0790-8_57

Computational Considerations 1896
7 Self-Similarity and Symmetry 1897
Rigid Symmetry 1897
Intrinsic Symmetry 1897
Spectral Symmetry 1897
Partial Symmetry 1898
Repeating Structure 1899
8 Feature-Based Methods 1899
Feature Descriptors 1899
Bags of Features 1901
Combining Global and Local Information 1901
9 Conclusion 1902
Cross-References 1903
References 1903

Abstract

Nonrigid shapes are ubiquitous in nature and are encountered at all levels of life, from macro to nano. The need to model such shapes and understand their behavior arises in many applications in imaging sciences, pattern recognition, computer vision, and computer graphics. Of particular importance is understanding which properties of the shape are attributed to deformations and which are invariant, i.e., remain unchanged. This chapter presents an approach to nonrigid shapes from the point of view of metric geometry. Modeling shapes as metric spaces, one can pose the problem of shape similarity as the similarity of metric spaces and harness tools from theoretical metric geometry for the computation of such a similarity.

1 Introduction

Those who played the game Rock, Paper, and Scissors in their childhood certainly remember the three gestures used in the game: Rock, represented by a clenched fist; Paper, represented by an open hand; and Scissors, represented by the extended index and middle fingers. These gestures are a toy example of the *nonrigid shape similarity* problem, which is the central topic of this chapter. No matter how one bends the fingers, he will immediately recognize the underlying object: the human hand.

More generally, the problem of determining the similarity of shapes undergoing certain class of transformations is termed *invariant shape similarity*. A similarity criterion is said to be invariant if it is not influenced by the transformation (Fig. 1). Different classes of transformations prescribe different similarity criteria based on geometric shape properties that are invariant under such transformations. The wider is the class, the less properties are preserved, and as a thumb rule, the more difficult is the problem. Specifically, in this chapter, we will consider rigid, inelastic, topology-changing, and scaling transformations. In many cases, such transformations are a good approximation of real transformations that natural objects may undergo.

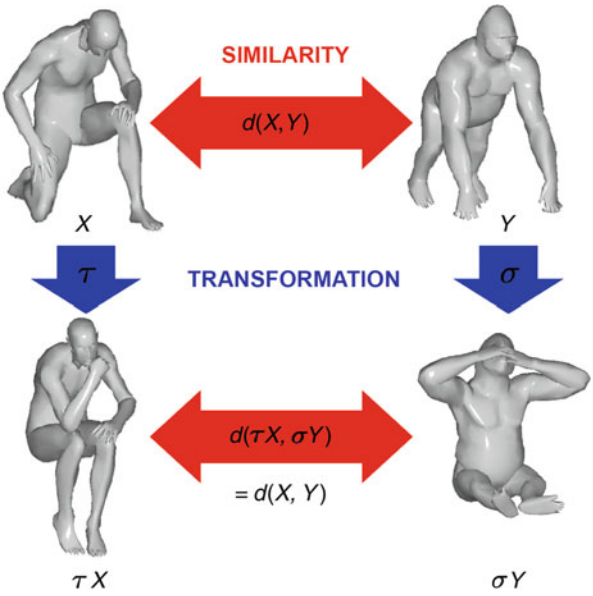


Fig. 1 Invariant shape similarity

Problems

Since nonrigid shapes are ubiquitous in the world and are encountered at all scales from macro to nano, nonrigid shape similarity plays a key role in many applications in imaging sciences, pattern recognition, computer vision, and computer graphics. Two archetype problems in shape analysis considered in this chapter are *invariant similarity* and *correspondence*. As will be discussed in the following, these two problems are interrelated: finding the best correspondence between two shapes also allows quantifying their similarity.

A good example of shape similarity is the problem of face recognition [19,21,24]. As the crudest approximation, one can think of faces as rigid surfaces and compare them using similarity criteria invariant under rigid transformations. However, such an approach does not account for surface deformations due to facial expressions, which can be approximated by inelastic deformations. Accounting for such deformations requires different similarity criteria. Yet, even elastic deformations are not enough to model the behavior of human faces: many facial expresses involve elastic deformations that change the facial shape topology (think of open and closed mouth). This extension of the model will require revisiting the similarity criterion once again.

The problem of correspondence is often encountered in shape synthesis applications such as morphing. In order to morph one shape into the other, one needs to know which point on the first shape will be transformed into a point on the second shape, in other words, establishing a correspondence between the shapes.

Methods

Many different approaches to shape similarity and correspondence can be considered as instances of the *minimum-distortion correspondence problem*, in which two shapes are endowed with certain structure, and one attempts to find the best (least distorting) matching between these structures. Such structures can be *local* (e.g., multiscale heat kernel signatures [106], local photometric properties [107, 119], or conformal factor [12]) or *global* (e.g., geodesic [24, 47, 78], diffusion [28], and commute time [31]) distances. The distortion of the best possible correspondence can be used as a criterion of shape similarity. By defining a structure invariant under certain class of transformations, it is possible to obtain invariant correspondence or similarity.

Local structures can be regarded as feature descriptors. As a model for global structures, metric spaces are used.

Chapter Outline

This chapter tries to present a unifying view on the archetypical problems in shape analysis. The first part presents a metric view on the problem of nonrigid shape similarity and correspondence, a common denominator allowing to deal with different types of invariance. According to this model, shapes are represented as metric spaces. The mathematical foundations of this model are provided in Sect. 2. Sections 3 and 4 deal with discrete representation of shapes, which is of essence in practical numerical computations. Section 5 provides a rigorous formulation of the invariant shape similarity problem and reviews different algorithms for its computation. Section 6 deals with an extension of invariant similarity to shapes which are partially similar, and Sect. 7 deals with a particular case of self-similarity and symmetry. Local feature-based methods and their use to create global shape descriptors are presented in Sect. 8. Finally, concluding remarks in Sect. 9 end the chapter. This chapter is based in part on the book [26], to which the reader is referred for further discussion and details.

2 Shapes as Metric Spaces

Elad and Kimmel [47, 48], Mémoli and Sapiro [78], and Bronstein et al. [23, 24] suggested to model shapes as metric spaces. The key idea of this model is that it allows to compare shapes as metric spaces. Since the model allows arbitrarity in the definition of the metric, desired invariance considerations guide the choice of the metric.

This section introduces the mathematical formalism and notation of this model and shows the construction of three different types of metric geometries: Euclidean, Riemannian, and diffusion.

Basic Notions

Topological Spaces

Given a set X, a *topology* T on X is a collection of subsets of X satisfying (Ti) $X, 0 \in T$; (Tii) $\bigcup_\alpha U_\alpha \in T$ for $U_\alpha \in T$; and (Tiii) $\bigcap_{i=1}^{N} U_i \in T$ for $U_i \in T$. X together with T is called a *topological space*. By convention, sets in T are referred to as *open sets* and their complements as *closed sets*.

A *neighborhood* $N(x)$ of x is a set containing an open set $U \in T$ such that $x \in U$. Points with neighborhood are called *interior*.

A topological space is called *Hausdorff* if distinct points in it have disjoint neighborhoods.

Two topological spaces X and Y are *homeomorphic* if there exists bijection $\alpha : X \to Y$ which is continuous and has continuous inverse α^{-1}. Since homeomorphisms copy topologies, homeomorphic spaces are topologically equivalent [1].

Metric Spaces

A function $d : X \times X \to \mathbb{R}$ which is (Mi) *positive-definite* ($d(x, y) > 0$ for all $x \neq y$ and $d(x, y) = 0$ for $x = y$) and (Mii) *subadditive* ($d(x, z) \leq d(x, y) + d(y, z)$ for all x, y, z) is called a *metric* on X. The metric is an abstraction of the notion of distance between pairs of points on X. Property (Mii) is called *triangle inequality* and generalizes the known fact: the sum of the lengths of two edges of a triangle is greater or equal to the length of the third edge. The pair (X, d) is called a *metric space*.

A metric induces topology through the definition of *open metric ball* $B_r(x) = \{x' \in X : d(x, x') < r\}$. A neighborhood of x in a metric space is a set containing a metric ball $B_r(x)$ [34].

Isometries

Given two metric spaces (X, d) and (Y, δ), the set $C \subset X \times Y$ of pairs such that for every $x \in X$ there exists at least one $y \in Y$ such that $(x, y) \in C$, and similarly, for every $y \in Y$, there exists an $x \in X$ such that $(x, y) \in C$ is called a *correspondence* between X and Y. Note that a correspondence C is not necessarily a function. The correspondence is called *bijective* if every point in X has a unique corresponding point in Y and vice versa.

The discrepancy of the metrics d and δ between the corresponding points is called the *distortion* of the correspondence,

$$\operatorname{dis}(C) = \sup_{(x,y),(x',y') \in C} \left| d(x, x') - \delta(y, y') \right|.$$

Metric spaces (X, d) and (Y, δ) are said to be ϵ-*isometric* if there exists a correspondence C with $\operatorname{dis}(C) \leq \epsilon$. Such a C is called an ϵ-*isometry*.

A particular case of a 0-isometry is called an *isometry*. In this case, the correspondence is a bijection and X and Y are called *isometric*.

Euclidean Geometry

Euclidean space $\mathbb{R}^m$ (hereinafter also denoted as $\mathbb{E}$) with the *Euclidean metric* $d_{\mathbb{E}}(x, x') = \|x - x'\|_2$ is the simplest example of a metric space. Given as a subset X of $\mathbb{E}$, we can measure the distances between points x and x' on X using the *restricted Euclidean metric*,

$$d_{\mathbb{E}}|_{X\times X}(x, x') = d_{\mathbb{E}}(x, x')$$

for all x, x' in X.

The restricted Euclidean metric $d_{\mathbb{E}}|_{X\times X}$ is invariant under Euclidean transformations of X, which include translation, rotation, and reflection in $\mathbb{E}$. In other words, X and its Euclidean transformation $i(X)$ are isometric in the sense of the Euclidean metric. Euclidean isometries are called *congruences*, and two subsets of $\mathbb{E}$ differing up to a Euclidean isometry are said to be *congruent*.

Riemannian Geometry

Manifolds

A Hausdorff space X which is locally homeomorphic to $\mathbb{R}^n$ (i.e., for every x in X, there exists a neighborhood U and a homeomorphism $\alpha : U \to \mathbb{R}^n$) is called an *n-manifold* or an *n-dimensional manifold*. The function α is called a *chart*. A collection of neighborhoods that cover X together with their charts is called an *atlas* on X. Given two charts α and β with overlapping domains U and V, the map $\beta\alpha^{-1} : \alpha(U \cap V) \to \beta(U \cap V)$ is called a *transition function*. A manifold whose transition functions are all differentiable is called a *differentiable manifold*. More generally a $\mathcal{C}^k$*-manifold* has all transition maps k-times continuously differentiable. A $\mathcal{C}^\infty$-manifold is called *smooth*.

A *manifold with boundary* is not a manifold in the strict sense of the above definition. Its *interior points* are locally homeomorphic to $\mathbb{R}^n$, and every point on the *boundary* ∂X is homeomorphic to $[0, \infty) \times \mathbb{R}^{n-1}$.

Of particular interest for the discussion in this chapter are two-dimensional ($n = 2$) manifolds, which model boundaries of physical objects in the world surrounding us. Such manifolds are also called *surfaces*. In the following, when referring to shapes and objects, the terms manifold, surface, and shape will be used synonymously.

Differential Structures

Locally, a manifold can be represented as a linear space, in the following way. Let $\alpha : U \to \mathbb{R}^n$ be a chart on a neighborhood of x and $\gamma : (-1, 1) :\to X$ be a differentiable curve passing through $x = \gamma(0)$. The derivative of the curve $\frac{d}{dt}(\alpha \circ \gamma)(0)$ is called a *tangent vector* at x. The set of all equivalence classes of tangent vectors at x forming an n-dimensional real vector space is called the *tangent space* $T_x X$ at x.

A family of inner products $\langle \cdot, \cdot \rangle_x : T_x X \times T_x X \to \mathbb{R}$ depending smoothly on x is called *Riemannian metric tensor*. A manifold X with a Riemannian metric tensor is called a *Riemannian manifold*.

The Riemannian metric allows to define local length structures and differential calculus on the manifold. Given a differentiable scalar-valued function $f : X \to \mathbb{R}$, the *exterior derivative* (*differential*) is a form $df = \langle \nabla f, \cdot \rangle$ on the tangent space TX. For a tangent vector $\mathbf{v} \in T_x X$, $df(x)\mathbf{v} = \langle \nabla f(x), \mathbf{v} \rangle_x$. ∇f is called the *gradient* of f at x and is a natural generalization of the notion of the gradient in vector spaces to manifolds. Similarly to the definition of Laplacian satisfying

$$\int_X \langle \nabla f, \nabla h \rangle_x d\mu(x) = \int_X h \Delta_X f d\mu(x)$$

for differentiable scalar-valued functions f and h, the operator Δ_X is called the *Laplace–Beltrami operator*, a generalization of the Laplacian. Here, μ denotes the measure associated with the n-dimensional *volume element* (*area element* for $n = 2$). The Laplace–Beltrami operator is (Li) symmetric ($\int_X h \Delta_X f d\mu(x) = \int_X f \Delta_X x d\mu(x)$), (Lii) of local action ($\Delta_X f(x)$ is independent of the value of $f(x')$ for $x' \neq x$), and (Liii) positive semi-definite ($\int_X f \Delta_X f d\mu(x) \geq 0$) (in many references, the Laplace–Beltrami is defined as a negative semi-definite operator) and (Liv) coincides with the Laplacian on Euclidean domains, such that $\Delta_X f = 0$ if f is a linear function and X is Euclidean.

Geodesics

Another important use of the Riemannian metric tensor is to measure the length of paths on the manifold. Given a continuously differentiable curve $\gamma : [a, b] \to X$, its length is given by

$$\ell(\gamma) = \int_a^b \langle \gamma'(t), \gamma'(t) \rangle_{\gamma(t)}^{1/2} dt.$$

For the set of all continuously differentiable curves $\Gamma(x, x')$ between the points x, x',

$$d_X(x, x') = \inf_{\gamma \in \Gamma(x, x')} \ell(\gamma) \tag{1}$$

defines a metric on X referred to as *length* or *geodesic metric*. If the manifold is compact, for any pair of points x and x', there exists a curve $\gamma \in \Gamma(x, x')$ called a *minimum geodesic* such that $\ell(\gamma) = d_X(x, x')$.

Embedded Manifolds

A particular realization of a Riemannian manifold called *embedded manifold* (or *embedded surface* for $n = 2$) is a smooth submanifold of $\mathbb{R}^m$ ($m > n$). In this case, the tangent space is an n-dimensional hyperplane in $\mathbb{R}^m$, and the Riemannian metric

is defined as the restriction of the Euclidean inner product to the tangent space, $\langle \cdot, \cdot \rangle_{\mathbb{R}^m}|_{TX}$.

The length of a curve $\gamma : [a, b] \to X \subset \mathbb{R}^m$ on an embedded manifold is expressed through the Euclidean metric,

$$\ell(\gamma) = \int_a^b \left(\langle \gamma'(t), \gamma'(t) \rangle_{\mathbb{R}^m}|_{T_{\gamma(t)}X} \right)^{1/2} dt, = \int_a^b \|\gamma'(t)\|_{\mathbb{R}^m} dt \tag{2}$$

and the geodesic metric d_X defined according to (1) is said to be *induced* by $d_{\mathbb{R}^m}$. (Repeating the process, one obtains that the metric induced by d_X is equal to d_X. For this reason, d_X is referred to as *intrinsic metric* [34].)

Though apparently embedded manifolds are a particular case of a more general notion of Riemannian manifolds, it appears that any Riemannian manifold can be realized as an embedded manifold. This is a consequence of the *Nash embedding theorem* [84], showing that a $C^k (k \geq 3)$ Riemannian manifold can be isometrically embedded in a Euclidean space of dimension $m = n^2 + 5n + 3$. In other words, any smooth Riemannian manifold can be defined as a metric space which is isometric to a smooth submanifold of a Euclidean space with the induced metric.

Rigidity

Riemannian manifolds do not have a unique realization as embedded manifolds. One obvious degree of freedom is the set of all Euclidean isometries: two congruent embedded manifolds are isometric and thus are realizations of the same Riemannian manifold. However, a Riemannian manifold may have two realizations which are isometric but incongruent. Such manifolds are called *nonrigid*. If, on the other hand, a manifold's only isometries are congruences, it is called *rigid*.

Diffusion Geometry

Another type of metric geometry arises from the analysis of heat propagation on manifolds. This geometry is called *diffusion* and is also intrinsic. We start by reviewing properties of diffusion operators.

Diffusion Operators

A function $k : X \times X \to \mathbb{R}$ is called a *diffusion kernel* if it satisfies the following properties: (Ki) *nonnegativity:* $k(x, x) \geq 0$; (Kii) *symmetry:* $k(x, y) = k(y, x)$; (Kiii) *positive-semidefiniteness:* for every bounded f,

$$\int \int k(x, y) f(x) f(y) d\mu(x) d\mu(y) \geq 0;$$

(Kiv) *square integrability:* $\int \int k^2(x, y) d\mu(x) d\mu(y) < \infty$; and (Kv) *conservation:* $\int k(x, y) d\mu\ (y) = 1$. The value of $k(x, y)$ can be interpreted as a transition probability from x to y by one step of a random walk on X.

Diffusion kernel defines a linear operator

$$\mathbf{K}f = \int k(x,y)f(y)d\mu(y), \tag{3}$$

which is known to be self-adjoint. Because of (Kiv), $\mathbf{K}$ has a finite Hilbert norm and therefore is compact. As the result, it admits a discrete eigendecomposition $\mathbf{K}\psi_i = \alpha_i\psi_i$ with some eigenfunctions $\{\psi_i\}_{i=0}^{\infty}$ and eigenvalues $\{\alpha_i\}_{i=0}^{\infty}$. $\alpha_i \geq 0$ by virtue of property (Kiii), and $\alpha_i \leq 1$ by virtue of (Kv) and consequence of the Perron–Frobenis theorem.

By the spectral theorem, the diffusion kernel can be presented as $k(x,y) = \sum_{i=0}^{\infty}\alpha_i\psi_i(x)\psi_i(y)$. Since $\{\psi_i\}_{i=1}^{\infty}$ form an orthonormal basis of $L^2(X)$,

$$\int\int k^2(x,y)d\mu(x)d\mu(y) = \sum_{i=0}^{\infty}\alpha_i^2, \tag{4}$$

a fact sometimes referred to as Parseval's theorem. Using these results, properties (Kiii–Kv) can be rewritten in the spectral form as $0 \leq \alpha_i \leq 1$ and $\sum_{i=0}^{\infty}\alpha_i^2 < \infty$.

An important property of diffusion operators is the fact that for every $t \geq 0$, the operator $\mathbf{K}^t$ is also a diffusion operator with the eigenbasis of $\mathbf{K}$ and corresponding eigenvalues $\{\alpha_i^t\}_{i=0}^{\infty}$. The kernel of $\mathbf{K}^t$ expresses the transition probability by random walk of t steps. This allows to define a scale space of kernels, $\{k_t(x,y)\}_{t\in T}$, with the scale parameter t.

There exist a large variety of possibilities to define a diffusion kernel and the related diffusion operator. Here, we restrict our attention to operators describing *heat diffusion*. Heat diffusion on surfaces is governed by the *heat equation*

$$\left(\Delta_X + \frac{\partial}{\partial t}\right)u(x,t) = 0; \qquad u(x,0) = u_0(x), \tag{5}$$

where $u(x,t)$ is the distribution of heat on the surface at point x in time t, u_0 is the initial heat distribution, and Δ_X is the positive-semidefinite Laplace–Beltrami operator, a generalization of the second-order Laplacian differential operator Δ to non-Euclidean domains. (If X has a boundary, *boundary conditions* should be added.)

On Euclidean domains ($X = \mathbb{R}^m$), the classical approach to the solution of the heat equation is by representing the solution as a product of temporal and spatial components. The spatial component is expressed in the Fourier domain, based on the observation that the Fourier basis is the eigenbasis of the Laplacian Δ, and the corresponding eigenvalues are the frequencies of the Fourier harmonics. A particular solution for a point initial heat distribution $u_0(x) = \delta(x-y)$ is called the *heat*

kernel $h_t(x-y) = \frac{1}{(4\pi t)^{m/2}} e^{-\|x-y\|^2/4t}$, which is shift invariant in the Euclidean case. A general solution for any initial condition u_0 is given by convolution $\mathbf{H}^t u_0 = \int_{\mathbb{R}^m} h_t(x-y)u_0(y)dy$, where $\mathbf{H}^t$ is referred to as *heat operator*.

In the non-Euclidean case, the eigenfunctions of the Laplace–Beltrami operator $\Delta_X \phi_i = \lambda_i \phi_i$ can be regarded as a "Fourier basis" and the eigenvalues given the "frequency" interpretation. The heat kernel is not shift invariant but can be expressed as $h_t(x,y) = \sum_{i=0}^{\infty} e^{-t\lambda_i}\phi_i(x)\phi_i(y)$.

It can be shown that the heat operator is related to the Laplace–Beltrami operator as $\mathbf{H}^t = e^{-t\Delta}$, and as a result, it has the same eigenfunctions ϕ_i and corresponding eigenvalues $e^{-t\lambda_i}$. It can be thus seen as a particular instance of a more general family of diffusion operators $\mathbf{K}$ diagonalized by the eigenbasis of the Laplace–Beltrami operator, namely, $\mathbf{K}$'s as defined in the previous section but restricted to have the eigenfunctions $\psi_i = \phi_i$. The corresponding diffusion kernels can be expressed as

$$k(x,y) = \sum_{i=0}^{\infty} K(\lambda_i)\phi_i(x)\phi_i(y), \tag{6}$$

where $K(\lambda)$ is some function such that $\alpha_i = K(\lambda_i)$ (in the case of $\mathbf{H}_t$, $K(\lambda) = e^{-t\lambda}$). Since the Laplace–Beltrami eigenvalues can be interpreted as frequency, $K(\lambda)$ can be thought of as the *transfer function* of a low-pass filter. Using this signal processing analogy, the kernel $k(x,y)$ can be interpreted as the point spread function at a point y, and the action of the diffusion operator $\mathbf{K}f$ on a function f on X can be thought of as the application of the point spread function by means of a non shift-invariant version of convolution. The transfer function of the diffusion operator $\mathbf{K}^t$ is $K^t(\lambda)$, which can be interpreted as multiple applications of the filter $K(\lambda)$. Such multiple applications decrease the effective bandwidth of the filter and, consequently, increase its effective support in space. Because of this duality, both $k(x,y)$ and $K(\lambda)$ are often referred to as diffusion kernels.

Diffusion Distances

Since a diffusion kernel $k(x,y)$ measures the degree of proximity between x and y, it can be used to define a metric

$$d^2(x,y) = \|k(x,\cdot) - k(y,\cdot)\|^2_{L^2(X)}, \tag{7}$$

on X, which was first constructed by Berard et al. in [*] and dubbed as the *diffusion distance* by Coifman and Lafon [40]. Another way to interpret the latter distance is by considering the embedding $\Psi : x \mapsto L^2(X)$ by which each point x on X is mapped to the function $\Psi(x) = k(x,\cdot)$. The embedding Ψ is an isometry between X equipped with diffusion distance and $L^2(X)$ equipped with the standard L^2 metric, since $d(x,y) = \|\Psi(x) - \Psi(y)\|_{L^2(X)}$. Because of spectral duality, the diffusion distance can also be written as

$$d^2(x, y) = \sum_{i=0}^{\infty} K^2(\lambda_i)(\phi_i(x) - \phi_i(y))^2. \tag{8}$$

Here as well we can define an isometric embedding $\Phi : x \mapsto \ell^2$ with $\Phi(x) = \{K(\lambda_i)\phi_i(x)\}_{i=0}^{\infty}$, termed as the *diffusion map* by Lafon. The diffusion distance can be cast as $d(x, y) = \|\Phi(x) - \Phi(y)\|_{\ell^2}$.

The same way a diffusion operator $\mathbf{K}^t$ defines a scale space, a family of diffusion metrics can be defined for $t \geq 0$ as

$$\begin{aligned} d_t^2(x, y) &= \|\Phi_t(x) - \Phi_t(y)\|_{\ell^2}^2 \\ &= \sum_{i=0}^{\infty} K^{2t}(\lambda_i)(\phi_i(x) - \phi_i(y))^2, \end{aligned} \tag{9}$$

where $\Phi_t(x) = \{K^t(\lambda_i)\phi_i(x)\}_{i=0}^{\infty}$. Interpreting diffusion processes as random walks, d_t can be related to the "connectivity" of points x and y by walks of length t (the more such walks exist, the smaller is the distance).

The described framework is very generic, leading to numerous potentially useful diffusion geometries parametrized by the selection of the transfer function $K(\lambda)$. Two particular choices are frequent in shape analysis, the first one being the *heat kernel*, $K_t(\lambda) = e^{-t\lambda}$, and the second one being the *commute time kernel*, $K(\lambda) = \frac{1}{\sqrt{\lambda}}$, resulting in the *heat diffusion* and *commute time* metrics, respectively. While the former kernel involves a scale parameter, typically tuned by hand, the latter one is *scale invariant*, meaning that neither the kernel nor the diffusion metric it induces changes under uniform scaling of the embedding coordinates of the shape.

3 Shape Discretization

In order to allow storage and processing of a shape by a digital computer, it has to be *discretized*. This section reviews different notions in the discrete representation of shapes.

Sampling

Sampling is the reduction of the continuous surface X representing a shape into a finite discrete set of representative points $\hat{X} = \{x_1, \ldots, x_N\}$. The number of points $|\hat{X}| = N$ is called the *size* of the sampling. The *radius* of the sampling refers to the smallest positive scalar r for which $\hat{X}$ is an r-covering of X, i.e.,

$$r(\hat{X}) = \max_{x \in X} \min_{x_i \in \hat{X}} d_X(x, x_i). \tag{10}$$

The sampling is called *s-separated* if $d_X(x_i, x_j) \geq s$ for every distinct $x_i, x_j \in \hat{X}$.

Sampling partitions the continuous surface into a set of disjoint regions,

$$V_i = \{x \in X : d_X(x, x_i) < d_X(x, x_j), x_{j \neq i} \in \hat{X}\}, \tag{11}$$

called the *Voronoi regions* [7] (Fig. 2). A Voronoi region V_i contains all the points on X that are closer to x_i than to any other x_j. That the sampling is said to induce a *Voronoi tessellation* (Unlike in the Euclidean case where every sampling induces a valid tessellation (cell complex), samplings of curved surfaces may result in Voronoi regions that are not valid cells, i.e., are not homeomorphic to a disk. In [66], Leibon and Letscher showed that an r-separated sampling of radius r with r smaller than $\frac{1}{5}$ of the convexity radius of the shape is guaranteed to induce a valid tessellation.) which we denote by $V(\hat{X}) = \{V_i, \ldots, V_n\}$.

Sampling can be regarded as a *quantization* process in which a point x on the continuous surface is represented by the closest x_i in the sampling [51]. Such a process can be expressed as a function mapping each V_i to the corresponding (Points on the boundary of the Voronoi regions are equidistant from at least two sample points and therefore can be mapped arbitrarily to any of them.) sample x_i. Intuitively, the smaller are the Voronoi regions, the better is the sampling. Sampling quality is quantified using an *error function*. For example,

$$\epsilon_\infty(\hat{X}) = \max_{x \in X} d_X(x, \hat{X}) = \max_{x \in X} \min_{x_i \in \hat{X}} d_X(x, x_i) \tag{12}$$

determines the maximum size of the Voronoi regions. If the shape is further equipped with a measure (e.g., the standard area measure), other error functions can be defined, e.g.,

$$\epsilon_p(\hat{X}) = \sum_i \int_{V_i} d_X^p(x, x_i) d\mu(x). \tag{13}$$

In what follows, we will show sampling procedures optimal or nearly optimal in terms of these criteria.

Farthest Point Sampling

Farthest point sampling (FPS) is a greedy procedure constructing a sequence of samplings $\hat{X}_1, \hat{X}_2, \ldots$. A sampling $\hat{X}_{N+1}$ is constructed from $\hat{X}_N$ by adding the *farthest point*

$$x_{N+1} = \arg\max_{x \in X} d_X(x, \hat{X}_N) = \arg\max_{x \in X} \min_{x_i \in \hat{X}_N} d_X(x, x_i). \tag{14}$$

The sequence $\{r_N\}$ of the sampling radii associated with $\{\hat{X}_N\}$ is nonincreasing, and, furthermore, each $\hat{X}_N$ is also r_N-separated. The starting point x_1 is usually picked up at random, and the stopping condition can be either the sampling size or radius.

Fig. 2 Voronoi decomposition of a surface with a non-Euclidean metric

Though FPS does not strictly minimize any of the error criteria defined in the previous section, in terms of ϵ_∞, it is no more than twice inferior to the optimal sampling of the same size [59]. In other words, for $\hat{X}$ produced using FPS,

$$\epsilon_\infty(\hat{X}) \leq 2 \min_{|\hat{X}'|=|\hat{X}|} \epsilon_\infty(\hat{X}'). \tag{15}$$

This result is remarkable, as finding the optimal sampling is known to be an NP-hard problem.

Centroidal Voronoi Sampling

In order for a sampling to be ϵ_2-optimal, each sample x_i has to minimize

$$\int_{V_i} d_X^2(x, x_i) d\mu(x). \tag{16}$$

A point minimizing the latter quantity is referred to as the *centroid* of V_i. Therefore, an ϵ_2-optimal sampling induces a so-called *centroidal Voronoi tessellation* (CVT), in which the centroid of each Voronoi region coincides with the sample point inducing it [45, 90]. Such a tessellation and the corresponding *centroidal Voronoi sampling* are generally not unique.

A numerical procedure for the computation of a CVT of a shape is known as the Lloyd–Max algorithm [68,73]. Given some initial sampling $\hat{X}^1$ of size N (produced, e.g., using FPS), the Voronoi tessellation induced by it is computed. The centroids of each Voronoi region are computed, yielding a new sampling $\hat{X}^2$ of size N.

The process is repeated iteratively until the change of $\hat{X}^k$ becomes insignificant. While producing high-quality samplings in practice, the Lloyd–Max procedure is guaranteed to converge only to a local minimum of ϵ_2. For computational aspects of CVTs on meshes, the reader is referred to [90].

Shape Representation

Once the shape is sampled, it has to be represented in a way allowing computation of discrete geometric quantities associate with it.

Simplicial Complexes

The simplest representation of a shape is obtained by considering the points of the sampling as points in the ambient Euclidean space. Such a representation is usually referred to as a *point cloud*. Points in the cloud are called *vertices* and denoted by $\mathbf{X} = \{x_1, \ldots, x_N\}$. The notion of a point cloud can be generalized using the formalism of simplical complexes. For our purpose, an abstract *k-simplex* is a set of cardinality $k + 1$. A subset of a simplex is called a *face*. A set K of simplices is said to be an abstract *simplical complex* if any face of $\sigma \in K$ is also in K, and the intersection of any two simplices $\sigma_1, \sigma_2 \in K$ is a face of both σ_1 and σ_2. A simplicial k-complex is a simplicial complex in which the largest dimension of any simplex is k. A simplicial k-complex is said to be *homogeneous* if every simplex of dimension less than k is the face of some k-simplex. A *topological realization* $\bar{K}$ of a simplicial complex K maps K to a simplicial complex in $\mathbb{R}^n$, in which vertices are identified with the canonical basis of $\mathbb{R}^n$, and each simplex in K is represented as the convex hull of the corresponding points $\{\mathbf{e}_i\}$. A *geometric realization* $\phi_{\mathbf{X}}(\bar{K})$ is a map of the simplicial complex $\bar{K}$ to $\mathbb{R}^3$ defined by associating the standard basis vectors $\mathbf{e}_i \in \mathbb{R}^n$ with the vertex positions x_i.

In this terminology, a point cloud is a simplicial 0-complex having a *discrete topology*. Introducing the notion of neighborhood, we can define a subset $E \subset \mathbf{X} \times \mathbf{X}$ of pairs of vertices that are *adjacent*. Pairs of adjacent vertices are called *edges*, and the simplicial 1-complex $\mathbf{X} \cup E$ has a *graph topology*, i.e., the set of vertices $\mathbf{X}$ forms an *undirected graph* with the set of edges E. A simplicial 2-complex consisting of vertices, edges, and triangular faces built upon triples of vertices and edges is called a *triangular mesh*. The mesh is called *topologically valid* if it is homeomorphic to the underlying continuous surface X. This usually implies that the mesh has to be a two manifold. A mesh is called *geometrically valid* if it does not contain self-intersecting triangles, which happens if and only if the geometric realization $\phi_{\mathbf{X}}(\bar{K})$ of the mesh is bijective. Consequently, any point x on a geometrically valid mesh can be uniquely represented as $\mathrm{x} = \varphi_{\mathbf{X}}(\mathbf{u})$. The vector $\mathbf{u}$ is called the *barycentric coordinates* of x and has at most three nonzero elements. If the point coincides with a vertex, $\mathbf{u}$ is a canonical basis vector; if the point lies on an edge, $\mathbf{u}$ has two nonzero elements; otherwise, $\mathbf{u}$ has three nonzero elements and x lies on a triangular face.

A particular way of constructing a triangular mesh stems from the Voronoi tessellation induced by the sampling. We define the simplicial 3-complex as

$$\mathbf{X} \cup \{(\mathrm{x}_i, \mathrm{x}_j) : \partial V_i \cap \partial V_j \neq \emptyset\} \cup \{(\mathrm{x}_i, \mathrm{x}_j, \mathrm{x}_k) : \partial V_i \cap \partial V_j \cap \partial V_k \neq \emptyset\}, \quad (17)$$

in which a pair of vertices spans an edge and a triple of vertices spans a face if the corresponding Voronoi regions are adjacent. A mesh defined in this way is called a *Delaunay mesh*. (Unlike in the Euclidean case where every sampling induces a valid Delaunay triangulation, an invalid Voronoi tessellation results in a topologically invalid Delaunay mesh. In [66], Leibon and Letscher showed that under the same conditions sufficient for the existence of a valid Voronoi tessellation, the Delaunay mesh is also topologically valid.)

Parametric Surfaces

Shapes homeomorphic to a disk can be parametrized using a single global chart, e.g., on the unit square, $\mathrm{x} : [0, 1]^2 \to \mathbb{R}^3$. (Manifolds with more complex topology can still be parametrized in this way by introducing cuts that open the shape into a topological disk.) Such surfaces are called *parametric* and can be sampled directly in the parametrization domain. For example, if the parametrization domain is sampled on a regular Cartesian grid, the shape can be represented as three $N \times N$ arrays of x, y, and z values. Such a completely regular structure is called a *geometry image* [56, 69, 99] and can be thought indeed as a three-channel image that can undergo standard image processing such as compression. Geometry images are ideally suitable for processing by vector and parallel hardware.

Implicit Surfaces

Another way of representing a shape is by considering the isosurfaces $\{x : \Phi(x) = 0\}$ of some function Φ defined on a region of $\mathbb{R}^3$. Such a representation is called *implicit*, and it often arises in medical imaging applications, where shapes are two-dimensional boundaries created by discontinuities in volumetric data. Implicit representation can be naturally processed using level-set-based algorithms, and it easily handles arbitrary topology. A disadvantage is the bigger amount of storage commonly required for such representations.

4 Metric Discretization

Next step in the discrete representation of shapes is the discretization of the metric.

Shortest Paths on Graphs

The most straightforward approach to metric discretization arises from considering the shape as a graph in which neighbor vertices are connected. A path in the graph between vertices x_i, x_j is an ordered set of connected edges

$$\Gamma(x_i, x_j) = \{(x_{i_1}, x_{i_2}), (x_{i_2}, x_{i_3}), \ldots, (x_{i_k}, x_{i_{k+1}})\} \subset E, \tag{18}$$

where $x_{i_1} = x_i$ and $x_{i_{k+1}} = x_j$. The length of path Γ is the sum of its constituent edge lengths,

$$L(\Gamma(x_i, x_j)) = \sum_{n=1}^{k} \|x_{i_k} - x_{i_{k+1}}\|. \tag{19}$$

A minimum geodesic in a graph is the shortest path between the vertices,

$$\Gamma^*(x_i, x_j) = \arg \min_{\Gamma(x_i, x_j)} L(\Gamma(x_i, x_j)). \tag{20}$$

We can use $d_L(x_i, x_j) = L(\Gamma^*(x_i, x_j))$ as a discrete approximation to the geodesic metric $d_X(x_i, x_j)$.

According to the *Bellman optimality principle* [10], given $\Gamma^*(x_i, x_j)$ a shortest path between x_i and x_j and x_k a point on the path, the sub-paths $\Gamma^*(x_i, x_k)$ and $\Gamma^*(x_k, x_j)$ are the shortest paths between x_i, x_k and x_k, x_j, respectively. The length of the shortest path in the graph can be thus expressed by the following recursive equation:

$$d_L(x_i, x_j) = \min_{x_k : (x_k, x_j) \in E} \left\{ d_L(x_i, x_k) + \|x_k - x_j\| \right\}. \tag{21}$$

Dijkstra's Algorithm

A famous algorithm for the solution of the recursion (21) was proposed by Dijkstra. Dijkstra's algorithm measures the *distance map* $d(x_k) = d_L(x_i, x_k)$ from the source vertex x_i to all the vertices in the graph.

Initialize $d(x_i) = 0$, $d(x_k) = \infty$ for all $k \neq i$; queue of unprocessed vertices $Q = \{x_1, \ldots, x_N\}$.
while Q is non-empty **do**
Find vertex with smallest value of d in the queue

$$x = \arg \min_{x \in Q} d(x)$$

for all unprocessed adjacent vertices $x' \in Q : (x, x') \in E$ **do**

$$d(x') = \min \left\{ d(x'), d(x) + \|x - x'\| \right\}$$

end for
Remove x from Q.
end while

Every vertex in Dijkstra's algorithm is processed exactly once; hence, Nn outer iterations are performed. Extraction of vertex with smallest d is straightforward with $\mathcal{O}(N)$ complexity and can be reduced to $\mathcal{O}(\log N)$ using efficient data structures such as *Fibonacci heap*. In the inner loop, updating adjacent vertices in our case, since the graph is sparsely connected, is $\mathcal{O}(1)$. The resulting overall complexity is $\mathcal{O}(N \log N)$.

Metrication Errors and Sampling Theorem

Unfortunately, the graph distance d_L is an inconsistent approximation of d_X, in the sense that d_L usually does not converge to d_X when the sampling becomes infinitely dense. This phenomenon is called *metrication error*, and the reason is that the graph induces a metric inconsistent with d_X (Fig. 3). While metrication errors make in general the use of d_L an approximation of d_X disadvantageous, the Bernstein–de Silva–Langford–Tenenbaum theorem [14] states that under certain conditions, the graph metric d_L can be made as close as desired to the geodesic metric d_X. The theorem is formulated as a bound of the form

$$1 - \lambda_1 \leq \frac{d_L}{d_X} \leq 1 + \lambda_2, \tag{22}$$

where λ_1, λ_2 depend on shape properties, sampling quality, and graph connectivity. In order for d_L to represent d_X accurately, the sampling must be sufficiently dense, length of edges in the graph bounded, and sufficiently close vertices must be connected, usually in a non-regular manner.

Fast Marching

Eikonal Equation

An alternative to computation of a discrete metric on a discretized surface is the discretization of the metric itself. The distance map $d(x) = d_X(x_0, x)$ (Fig. 4) on the manifold can be associated with the time of arrival of a propagating front traveling with unit speed (illustratively, imagine a fire starting at point x_0 at time $t = 0$ and propagating from the source). Such a propagation obeys the *Fermat principle of least action* (the propagating front chooses the quickest path to travel, which coincides with the definition of the geodesic distance) and is governed by the *eikonal equation*

$$\|\nabla_X d\|_2 = 1, \tag{23}$$

where ∇_X is the intrinsic gradient on the surface X. Eikonal equation is a hyperbolic PDE with boundary conditions $d(x_0) = 0$; minimum geodesics are its characteristics. Propagation direction is the direction of the steepest increase of d and is perpendicular to geodesics.

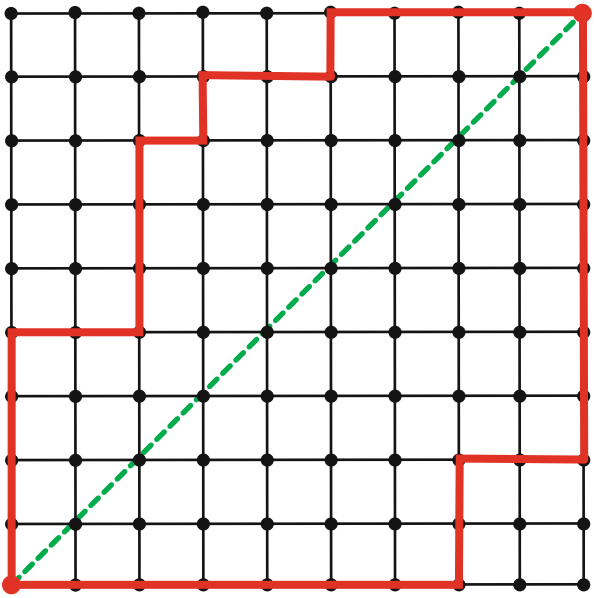


Fig. 3 Shortest paths measured by Dijkstra's algorithm (*solid bold lines*) do not converge to the true shortest path (*dashed diagonal*), no matter how much the grid is refined (Reproduced from [25])

Since the distance map is not everywhere differentiable (in particular, at the source point), no solution to the eikonal equation exists in the classical sense, while there exist many non $\mathcal{C}^1$ functions satisfying the equation and the boundary conditions. Among such functions, the largest d satisfying the boundary conditions and the inequality

$$\|\nabla_X d\|_2 \leq 1 \tag{24}$$

at every point where $\nabla_X d$ exists is called the *viscosity solution* [42]. The viscosity solution of the eikonal equation always exists and is unique, and its value at a point x coincides with $d_X(x, x_0)$. It is known to be monotonous, i.e., not having local maxima.

Triangular Meshes

A family of algorithms for finding the viscosity solution of the discretized eikonal equation by simulated wavefront propagation is called *fast marching methods* [63, 100, 112]. Fast marching algorithms can be thought of as continuous variants of Dijkstra's algorithm, with the notable difference that they consistently approximate the geodesic metric d_X on the surface.

Fig. 4 Distance map measured on a curved surface. Equidistant contours from the source located at the right hand are shown (Reproduced from [25])

Initialize $d(x_0) = 0$ and mark x_0 as *processed*; for all $k \neq 0$ set $d(x_k) = \infty$ and mark x_k as *unprocessed*.
while there exist *unprocessed* vertices **do**
  Mark *unprocessed* neighbors of *processed* vertices as *interface*.
  for all *interface* vertices x and all incident triangles (x, x_1, x_2) with $x_1, x_2 \neq$ *unprocessed* **do**
    Update $d(x)$ from $d(x_1)$ and $d(x_2)$.
  end for
  Mark *interface* vertex with the smallest value of d as *processed*.
end while

The general structure of fast marching closely resembles that of Dijkstra's algorithm with the main difference lying in the update step. Unlike the graph case where shortest paths are restricted to pass through the graph edges, the continuous approximation allows paths passing anywhere in the simplicial complex. For that reason, the value of $d(x)$ has to be computed from the values of the distance map at two other vertices forming a triangle with x. In order to guarantee consistency of the solution, all such triangles must have an acute angle at x. Obtuse triangles are split at a preprocessing stage by adding virtual connections to nonadjacent vertices.

Given a triangle (x, x_1, x_2) with known values of $d(x_1)$ and $d(x_2)$, the goal of the update step is to compute $d(x)$. The majority of fast marching algorithms do so by simulating the propagation of a planar wavefront in the triangle. The wavefront arrival time to x_1 and x_2 is set to $d(x_1)$ and $d(x_2)$, from which the parameters of the wave source are estimated. Generally, there exist two solutions for $d(x)$ consistent with the input, the smallest corresponding to the wavefront first arriving to x and

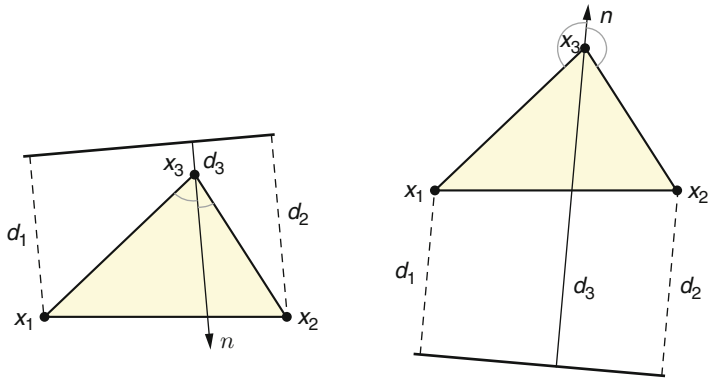


Fig. 5 Fast marching updates the triangle (x_1, x_2, x_3) by estimating the planar wavefront direction n and origin p based on d_1 at x_1 and d_2 at x_2 and propagating it further to x_3. d_3 has two possible solutions: the one shown on the *left* is inconsistent, since $d_3 < d_1, d_2$. The solution on the *right* is consistent, since $d_3 > d_1, d_2$. Geometrically, in order to be consistent, the update direction n has to form obtuse angles with the triangle edges (x_3, x_1) and (x_3, x_2) (Reproduced from [25])

then to x_1 and x_2 and the largest corresponding to the inverse situation. In order to guarantee monotonicity of the solution, the largest solution is always chosen (Fig. 5).

Computationally, fast marching has the $\mathcal{O}(N \log N)$ complexity of Dijkstra, perhaps with a slightly larger constant.

Parametric Surfaces

For surfaces admitting a global parametrization $x : U \to \mathbb{R}^3$, the eikonal equation can be expressed entirely in the parametrization domain as [104]

$$\nabla^{\mathrm{T}} d G^{-1} \nabla d = 1, \tag{25}$$

where $d(u)$ is the distance map in the parametrization domain, ∇d is its gradient with respect to the standard basis in $\mathbb{R}^2$, and G are the coefficients of the first fundamental form in parametrization coordinates. The fast marching update step can be therefore performed on U. Moreover, since only G is involved in the equation, the knowledge of the actual vertex coordinates is not required. This property is useful when the surface is reconstructed from some indirect measurements, e.g., normals or gradients, as it allows to avoid surface reconstruction for metric computation.

Parallel Marching

The main disadvantage of all Dijkstra-type algorithms based on a heap structure in general and fast marching in particular is the fact that they are inherently sequential. Moreover, as the order in which the vertices are visited is unknown in advance, they typically suffer from inefficient access to memory. Working with well-structured parametric surfaces such as geometry images allows to circumvent these

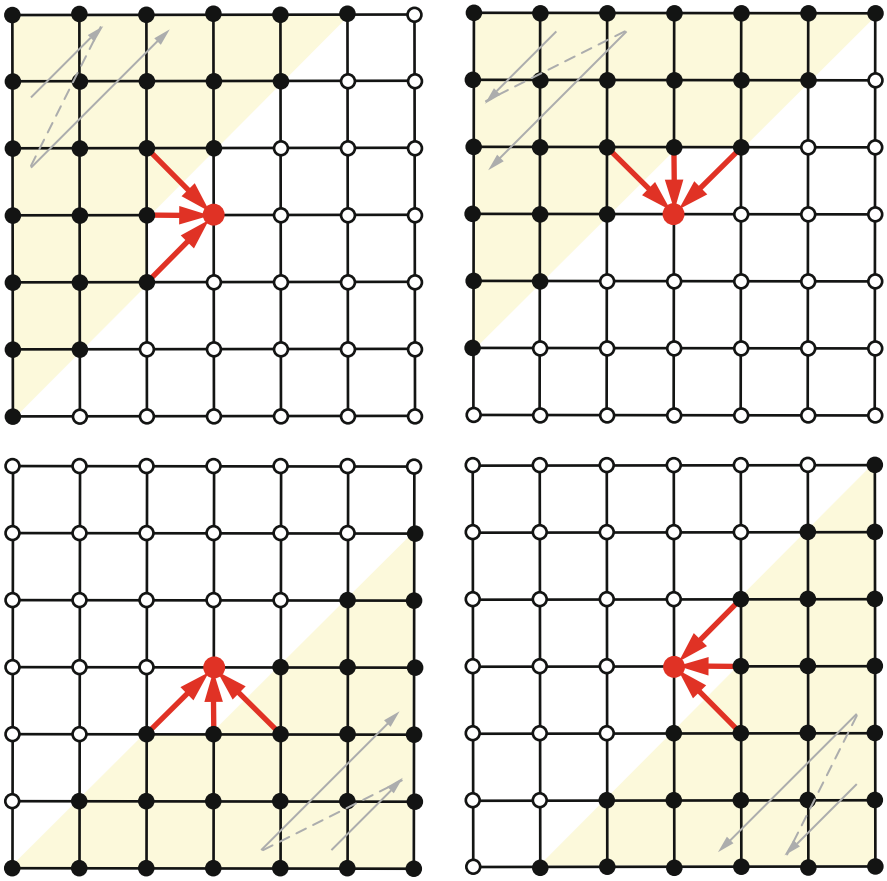


Fig. 6 Raster scan grid traversals rotated by 45° (Reproduced from [116])

disadvantages by replacing the heap-based update by the regular raster scan update. Such family of algorithms is usually called *parallel marching* [117] or *fast sweeping* [111, 121].

In parallel marching, vertices of the geometry image are visited in a raster scan order, and for each vertex, the standard fast marching update is applied using already updated (causal) vertices as the supporting vertices for the following update. Four raster scans in alternating left-to-right top-to-bottom, right-to-left top-to-bottom, left-to-right bottom-to-top, and right-to-left bottom-to-top directions are applied (in practice, it is advantageous to rotate the scan directions by 45°, as shown in Fig. 6). For a Euclidean domain, such four scans are sufficient to consistently approximate the metric; for non-Euclidean shapes, several repetitions of the four scans are required. The algorithm stops when the distance map stops changing significantly from one repetition to another. The exact number of repetitions required depends on the metric and the parametrization, but is practically very small.

Parallel marching algorithms map well on modern vector and parallel architectures and in particular on graphics hardware [117].

Implicit Surfaces and Point Clouds

Two-dimensional manifolds represented in the implicit form $X = \{\Phi(x) = 0\} \subset \mathbb{R}^3$ can be approximated with arbitrary precision as the union of Euclidean balls of radius $h > 0$ around X,

$$B_h(X) = \bigcup_{x \in X} B_h^{\mathbb{R}^3}(x). \tag{26}$$

$B_h(X)$ is a three-dimensional Euclidean submanifold, which for $h < 1/\max \kappa_2$ has a smooth boundary. For every $x, x' \in X$, the shortest path in $B_h(X)$ is no longer than the corresponding shortest path on X. Mémoli and Sapiro [77] showed that as $h \to 0$, shortest paths in $B_h(X)$ converge to those on X and the corresponding geodesic distances $d_{B_h(X)}|_{X \times X}$ converge uniformly to d_X. This result allows to cast the computation of a distance map on a curved two-dimensional space as the computation of a distance map on a three-dimensional Euclidean submanifold. The latter can be done using fast marching or parallel marching on orthogonal grid restricted to a *narrow band* around X [76].

A similar methodology can be used for the computation of distance maps on point clouds [76]. The union of Euclidean balls centered at each vertex of the cloud creates a three-dimensional Euclidean manifold, on which the distance map is computed using fast marching or parallel marching.

Diffusion Distance

The diffusion distance is expressed through the spectral decomposition of the Laplace–Beltrami operator, and its discretization involves the discretization of the Laplace–Beltrami operator and the computation of its eigenfunctions.

Discretized Laplace–Beltrami Operator

A discrete approximation of the Laplace–Beltrami on the mesh $\hat{X}$ has the following generic form

$$(\Delta_{\hat{X}} f)_i = \frac{1}{a_i} \sum_j w_{ij} (f_i - f_j), \tag{27}$$

where $f = (f_1, \ldots, f_N)$ is a scalar function defined on the mesh $\hat{X}$, w_{ij} are weights, and a_i are normalization coefficients. In matrix notation, Eq. (27) can be written as

$$\Delta_{\hat{X}} f = A^{-1} L f, \tag{28}$$

where $A = \mathrm{diag}(a_i)$ and $L = \mathrm{diag}\left(\sum_{l \neq i} w_{il}\right) - (w_{ij})$.

Different discretizations of the Laplace–Beltrami operator lead to different choice of A and W. In general, it is common to distinguish between *discrete* and *discretized* Laplace–Beltrami operator, the former being a combinatorial construction and the latter a discretization trying to preserve some of the properties (Li)–(Liv) of the continuous counterpart. In addition to these properties, it is important that the discrete Laplace–Beltrami operator converges to the continuous one, in the sense that the solution of the continuous heat equation with Δ_X converges to the discrete solution of the discrete heat equation with $\Delta_{\hat{X}}$ as the number of samples grows to infinity.

Purely combinatorial approximations such as the *umbrella operator* ($w_{ij} = 1$ if x_i and x_j are connected by an edge and zero otherwise) [120] and the *Tutte Laplacian* ($w_{ij} = d_i^{-1}$, where d_i is the valence of vertex x_i) [113] are not geometric, violate property (Liv), and do not converge to the continuous Laplace–Beltrami operator. One of the most widely used discretizations is the *cotangent weight* scheme [91] and its variants [79] ($w_{ij} = \cot \alpha_{ij} + \cot \beta_{ij}$ if x_i and x_j are connected, where α_{ij} and β_{ij} are the two angles opposite to the edge between vertices x_i and x_j in the two triangles sharing the edge and a_i is proportional to the sum of the areas of the triangles sharing x_i). It preserves properties (Li)–(Liv) as well as satisfies the convergence property under certain mild conditions [116].

Computation of Eigenfunctions and Eigenvalues

By solving the *generalized eigendecomposition* problem [67],

$$A\Phi = \Lambda L \Phi,$$

where Φ is an $N \times (k+1)$ matrix whose columns are discretized eigenfunctions $\phi_0, \ldots, \phi_k$ and Λ is the diagonal matrix of the corresponding eigenvalues $\lambda_0, \ldots, \lambda_k$ of the discretized Laplace–Beltrami operator are computed. ϕ_{il} approximates the value of the lth eigenfunction at the point x_i.

A different approach to the computation of eigenfunction is based on the *finite element method* (FEM). Using the Green formula, the Laplace–Beltrami eigenvalue problem $\Delta_X \phi = \lambda \phi$ can be expressed in the *weak form* as

$$\langle \Delta_X \phi, \alpha \rangle_{L_2(X)} = \lambda \langle \phi, \alpha \rangle_{L_2(X)} \tag{29}$$

for any smooth α. Given a finite basis $\{\alpha_1, \ldots, \alpha_K\}$ spanning a subspace of $L_2(X)$, the solution ϕ can be expanded as $\phi(x) \approx u_1 \alpha_1(x) + \cdots + u_K \alpha_K(x)$. Substituting this expansion into (29) results in a system of equations

$$\sum_{j=1}^{K} u_j \langle \Delta_X \alpha_j, \alpha_k \rangle_{L_2(X)} = \lambda \sum_{j=1}^{K} u_j \langle \alpha_j, \alpha_k \rangle_{L_2(X)}, \quad k = 1, \ldots, K,$$

which, in turn, is posed as a generalized eigenvalue problem

$$Au = \lambda Bu. \tag{30}$$

(here A and B are $K \times K$ matrices with elements $a_{kj} = \langle \Delta_X \alpha_j, \alpha_k \rangle_{L_2(X)}$ and $b_{kj} = \langle \alpha_j, \alpha_k \rangle_{L_2(X)}$). Solution of (30) gives eigenvalues λ and eigenfunctions $\phi = u_1\alpha_1 + \cdots + u_K\alpha_K$ of Δ_X.

As the basis, linear, quadratic, or cubic polynomials defined on the mesh can be used. Since the inner products are computed on the surface, the method is less sensitive to shape discretization compared to the direct approach based on the discretization of the Laplace–Beltrami operator. This is confirmed by numerical studies performed by Reuter et al. [94].

Discretization of Diffusion Distances

Using the discretized eigenfunctions, a discrete diffusion kernel is approximated as

$$K(x_i, x_j) \approx \sum_{l=0}^{k} K(\lambda_l)\phi_{il}\phi_{jl},$$

and can be represented as an $N \times N$ matrix. The corresponding diffusion distance is approximated as

$$d_{X,t}(x_i, x_j) \approx \left(\sum_{l=1}^{k} K^2(\lambda_i)(\phi_{il} - \phi_{jl})^2 \right)^{1/2}.$$

5 Invariant Shape Similarity

Let us denote by $\mathbb{X}$ the space of all shapes equipped with some metric, i.e., a point in $\mathbb{X}$ is a metric space (X, d_X). Let $\mathcal{T}$ be a group of shape *transformations*, i.e., a collection of operators $\tau : \mathbb{X} \to \mathbb{X}$ with the function composition. Two shapes differing by a transformation $\tau \in \mathcal{T}$ are said to be *equivalent up to* $\mathcal{T}$. The equivalence relation induces the *quotient space* $\mathbb{X}/\mathcal{T}$ in which each point is an *equivalence class* of shapes that differ by a transformation in $\mathcal{T}$. A particular instance of $\mathcal{T}$ is the group of *isometries*, i.e., such transformations that acting on X leave d_X unchanged. The exact structure of such the isometry group depends on the choice of the metric with which the shapes are equipped. For example, if the Euclidean metric $d_X = d_{\mathbb{E}}|_{X\times X}$ is used, the isometry group coincides with the group of Euclidean congruences (rotations, translations, and reflections).

A function $d : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ that associates a pair of shapes with a nonnegative scalar is called a *distance* or *dissimilarity* function. We will say that the dissimilarity d is $\mathcal{T}$*-invariant* if it defines a *metric* on the quotient space $\mathbb{X}/\mathcal{T}$. In particular,

this means that $d(X, \tau(X)) = 0$ and $d(\tau(X) \times \sigma(Y)) = d(X, Y)$ for every $\tau, \sigma \in \mathcal{T}$ and $X, Y \in \mathbb{X}$. In particular, for $\mathcal{T}$ being the isometry group, a $\mathcal{T}$-invariant dissimilarity is an *isometry-invariant* metric between shapes. The exact type of invariance depends on the structure of the isometry group and, hence, again on the choice of the metric with which the shapes are equipped.

As a consequence from the metric axioms, an isometry-invariant dissimilarity $d(X, Y)$ between two shapes X and Y equals to zeros if and only if X and Y are isometric. However, since exact isometry is an ideal rather than practical notion, it is desirable to extend this property to *similar* (almost isometric) rather than strictly equivalent (isometric) shapes. We will therefore require that (Ii) two ϵ-isometric shapes X and Y satisfy $d(X, Y) \leq c_1\epsilon + b_1$, and vice versa (Iii) if $d(X, Y) \leq \epsilon$, then X and Y are $(c_2\epsilon + b_2)$-isometric, where c_1, c_2, b_1, and b_2 are some nonnegative constants. In what follows, we will focus on the construction of such dissimilarities and their approximation and show how different choices of the metric yield different classes of invariance.

Rigid Similarity

Equipping shapes with the restriction of the Euclidean metric in $\mathbb{E}$ allows to consider them as subsets of a bigger common metric space, $\mathbb{E}$ equipped with the standard Euclidean metric. We will therefore examine dissimilarity functions allowing to compare between two subsets of the same metric space.

Hausdorff Distance

For two sets X and Y, a subset $R \subseteq X \times Y$ is said to be a *correspondence between X and Y* if (1) for every $x \in X$ there exists at least one $y \in Y$ such that $(x, y) \in R$ and (2) for every $y \in Y$ there exists at least one $x \in X$ such that $(x, y) \in R$. We will denote by $\mathcal{R}(X, Y)$ the set of all possible correspondences between X and Y.

Using the notion of correspondence, we can define the *Hausdorff distance* [58] between the two subsets of some metric space $(\mathbb{Z}, d_{\mathbb{Z}})$ as

$$d_{\mathrm{H}}^{\mathbb{Z}}(X, Y) = \min_{R \in \mathcal{R}(X,Y)} \max_{(x,y) \in R} d_{\mathbb{Z}}(x, y). \tag{31}$$

In other words, Hausdorff distance is the smallest nonnegative radius r for which $B_r(X) = \bigcup_{x \in X} B_r(x) \subseteq Y$ and $B_r(Y) \subseteq X$, i.e.,

$$d_{\mathrm{H}}^{\mathbb{Z}}(X, Y) = \max \left\{ \max_{x \in X} \min_{y \in Y} d_{\mathbb{Z}}(x, y), \max_{y \in Y} \min_{x \in X} d_{\mathbb{Z}}(x, y) \right\}. \tag{32}$$

Hausdorff distance is a metric on the set of all compact nonempty sets in $\mathbb{Z}$. However, it is not isometry invariant, i.e., for a nontrivial $\tau \in \mathrm{Iso}(\mathbb{Z})$, generally $d_{\mathrm{H}}^{\mathbb{Z}}(X, \tau(X)) \neq 0$. The isometry-invariant Hausdorff metric is constructed as the minimum of $d_{\mathrm{H}}^{\mathbb{Z}}$ over all isometries in $\mathbb{Z}$,

$$d_{\mathrm{H}}^{\mathbb{Z}/\mathrm{Iso}(\mathbb{Z})}(X,Y) = \min_{\tau\in\mathrm{Iso}(\mathbb{Z})} d_{\mathrm{H}}^{\mathbb{Z}}(X,\tau(Y)). \tag{33}$$

In the particular case of $(\mathbb{Z}, d_{\mathbb{Z}})$ being $(\mathbb{E}, d_{\mathbb{E}})$, the isometry-invariant Hausdorff metric can be used to quantify similarity between rigid shapes measuring to which extent they are *congruent* (isometric in the Euclidean sense) to each other. The metric assumes the form

$$d_{\mathrm{H}}^{\mathbb{E}/\mathrm{Iso}(\mathbb{E})}(X,Y) = \min_{R,t} d_{\mathrm{H}}^{\mathbb{E}}(X, RY + t), \tag{34}$$

where an orthogonal (rotation or, sometimes, rotation and reflection) matrix R and a translation vector $t \in \mathbb{E}$ are used to parametrize the Euclidean isometry group.

Iterative Closest Point Algorithms

Denoting by

$$\mathrm{cp}_Y(x) = \min_{y\in Y} d_{\mathbb{E}}(x,y) \tag{35}$$

the *closest point* to x in Y, the Euclidean isometry-invariant Hausdorff metric can be expressed as

$$\begin{aligned} d_{\mathrm{H}}^{\mathbb{E}/\mathrm{Iso}(\mathbb{E})}(X,Y) &= \max\left\{\min_{R,t}\max_{x\in X} d_{\mathbb{E}}(x, RY+t), \min_{R,t}\max_{y\in Y} d_{\mathbb{E}}(y, RX+t)\right\} \\ &= \min_{R,t}\max\left\{\max_{x\in X}\|x - \mathrm{cp}_{RY+t}(x)\|_2, \max_{y\in Y}\|y - \mathrm{cp}_{R^{-1}(X-t)}(y)\|_2\right\}. \end{aligned} \tag{36}$$

Such a formulation lends itself to numerical computation. A family of algorithms referred to as *iterative closest point* (ICP) [15, 36] first established the closest point correspondences between X and Y; once the correspondence is available, the Euclidean isometry (R,t) minimizing $\max_{x\in X}\|x-\mathrm{cp}_{RY+t}(x)\|_2$ and $\max_{y\in Y}\|y-\mathrm{cp}_{R^{-1}(X-t)}(y)\|_2$ is found and applied to Y. This, however, is likely to change the correspondence, so the process is repeated until convergence. For practical reasons, more robust variants of the Hausdorff distance are used [81].

Shape Distributions

A disadvantage of the ICP algorithms is that the underlying optimization problem becomes computationally intractable in high-dimensional spaces. A different approach for isometry-invariant comparison of rigid shapes, proposed by Osada et al. [85] and referred to as *shape distribution*, compares the distributions (histograms) of distances defined on the shape. Two isometric shapes obviously have identical shape distributions, which makes the approach isometry invariant. Shape distributions can be computed in a space of any dimension, are computationally efficient, and are not limited to a specific metric. A notable disadvantage of shape distribution

distance is that it does not satisfy our axioms (Ii)–(Iii), as there may be two non-isometric shapes with equal shape distributions; therefore, it is not a metric.

Wasserstein Distances

Let the sets X and Y be further equipped with measures μ_X and μ_Y, respectively. (It is required that $\mathrm{supp}(\mu_X) = X$ and $\mathrm{supp}(\mu_Y) = Y$.) We will say that a measure μ on $X \times Y$ is a coupling of μ_X and μ_Y if (i) $\mu(X' \times Y) = \mu_X(X')$ and (ii) $\mu(X \times Y') = \mu_Y(Y')$ for all Borel sets $X' \subseteq X$ and $Y' \subseteq Y$. We will denote by $\mathcal{M}(\mu_X, \mu_Y)$ the set of all possible couplings of μ_X and μ_Y. The *support* $\mathrm{supp}(\mu)$ of the measure μ is the minimum closed subset $R \subset X \times Y$ such that $\mu(R^c) = 0$. The support of each $\mu \in \mathcal{M}(\mu_X, \mu_Y)$ defines a correspondence; measure coupling can be therefore interpreted as a "soft" or "fuzzy" correspondence between two sets.

The family of distances

$$d^{\mathbb{Z}}_{\mathrm{W},p}(\mu_X, \mu_Y) = \min_{\mu \in \mathcal{M}(\mu_X, \mu_Y)} \left(\int_{X \times Y} d^p_{\mathbb{Z}}(x, y) d\mu(x, y) \right)^{\frac{1}{p}} \tag{37}$$

for $1 \leq p < \infty$, and

$$d^{\mathbb{Z}}_{\mathrm{W},\infty}(\mu_X, \mu_Y) = \min_{\mu \in \mathcal{M}(\mu_X, \mu_Y)} \max_{(x,y) \in \mathrm{supp}(\mu)} d_{\mathbb{Z}}(x, y) \tag{38}$$

for $p = \infty$ is called the *Wasserstein* or *Earth mover's distances* [97]. Wasserstein distances are metrics on the space of distributions (finite measures) on $\mathbb{Z}$. For convenience, we will sometimes write $d^{\mathbb{Z}}_{\mathrm{W},p}(X, Y)$ implying $d^{\mathbb{Z}}_{\mathrm{W},p}(\mu_X, \mu_Y)$.

Exactly like in the case of the Hausdorff distance, Wasserstein distances can be transformed into isometry-invariant metrics by considering the quotient with all isometries of $\mathbb{Z}$,

$$d^{\mathbb{Z}/\mathrm{Iso}(\mathbb{Z})}_{\mathrm{W},p}(X, Y) = \min_{\tau \in \mathrm{Iso}(\mathbb{Z})} d^{\mathbb{Z}}_{\mathrm{W},p}(X, \tau(Y)). \tag{39}$$

Wasserstein distances are intimately related to *Monge–Kantorovich optimal transportation problems*. Informally, if the measures μ_X and μ_Y are interpreted as two ways of piling up a certain amount of dirt over the regions X and Y, respectively, and the cost of transporting dirt from point x to point y is quantified by $d^p_{\mathbb{Z}}(x, y)$, then the Wasserstein distance $d^{\mathbb{Z}}_{\mathrm{W},p}$ expresses the minimum cost of turning one pile into the other. On discrete domains, the Wasserstein distance can be cast as an optimal *assignment problem* and solved using the *Hungarian algorithm* or *linear programming* [97]. Several approximations have also been proposed in [60, 102].

Canonical Forms

The Hausdorff distance allows comparing shapes equipped with the restricted Euclidean metric, i.e., considered as subsets of the Euclidean space. If other metrics

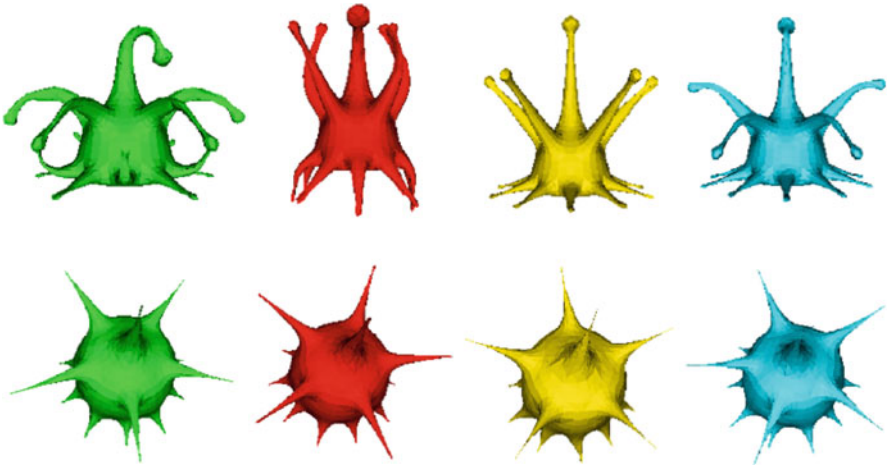


Fig. 7 Nearly isometric deformations of a shape (*top row*) and their canonical forms in $\mathbb{R}^3$ (*bottom row*)

are used, we have a more difficult problem of comparison of two different metric spaces. Elad and Kimmel [47,48] proposed an approximate solution to this problem, reducing it to the comparison of Euclidean subspaces by means of *minimum-distortion embedding*. Given a shape X with some metric d (e.g., geodesic or diffusion), it can be represented as a subset of the Euclidean space by means of an *embedding* $\varphi : X \to \mathbb{R}^m$. If the embedding is isometric ($d_{\mathbb{E}}|_{\varphi(X)\times\varphi(X)} \circ \varphi = d$), the Euclidean representation $(\varphi(X), d_{\mathbb{E}}|_{\varphi(X)\times\varphi(X)})$ called the *canonical form* of X can be used equivalently instead of (X, d) (Fig. 7). Given a Euclidean isometry $i \in \mathbb{E}$, if φ is isometric, then $\varphi \circ i$ is also isometric. In other words, the canonical form is defined up to an isometry. In a more general setting, an arbitrary metric space $(\mathbb{Z}, d_{\mathbb{Z}})$ is used instead of $\mathbb{E}$ for the computation of the canonical form.

The advantage of using canonical forms is that it brings the problem of shape comparison to the comparison of subsets of the Euclidean space, using, e.g., the Hausdorff distance. Given two shapes (X, d) and (Y, δ), their canonical forms $\varphi(X)$ and $\psi(Y)$ in $\mathbb{Z}$ are computed. In order to compensate for ambiguity in the definition of the canonical forms, an isometry-invariance distance between subsets of $\mathbb{Z}$ must be used, e.g., $d_{\mathrm{H}}^{\mathbb{Z}/\mathrm{Iso}(\mathbb{Z})}(\varphi(X), \psi(Y))$. In the particular case of Euclidean canonical forms, $d_{\mathrm{H}}^{\mathbb{E}/\mathrm{Iso}(\mathbb{E})}(\varphi(X), \psi(Y))$ can be computed using ICP.

Multidimensional Scaling

Unfortunately, in most cases, there exists no isometric embedding of X into some predefined metric space. The right choice of $\mathbb{Z}$ can decrease the embedding distortion, but cannot make it zero [21, 114]. Instead, one can find an embedding with minimal distortion,

$$\min_{\varphi:(X,d)\to(\mathbb{Z},d_{\mathbb{Z}})} \mathrm{dis}(\varphi).$$

In this case, $d_{\mathbb{Z}}|_{\varphi(X)\times\varphi(X)} \circ \varphi \approx d$, and thus the canonical form is only an approximate representation of the shape X in the space $\mathbb{Z}$.

In the discrete setting and $\mathbb{Z} = \mathbb{R}^m$, given the discretized shape $\{x_1, \ldots, x_N\}$ with the discretized metric $d_{ij} = d(x_i, x_j)$, the minimum-distortion embedding can be computed by solving the *multidimensional scaling* (MDS) problem [17,41],

$$\min_{\{z_1,\ldots,z_N\}\subset\mathbb{R}^m} \max_{i,j=1,\ldots,N} |d_{ij} - \|z_i - z_j\||, \tag{40}$$

where $z_i = \varphi(x_i)$.

In practical applications, other norms (e.g., L_2) are used in the MDS problem (40). The MDS objective function is usually referred to as *stress* in MDS literature. For the L_2 MDS problem (also known as *least squares* or LS-MDS), an efficient algorithm based on *iterative majorization* (commonly referred to as *scaling by majorizing a complicated function* or SMACOF) exists [44]. Denoting by Z the $N \times m$ matrix of the embedding coordinates of $\{x_1, \ldots, x_N\}$ in $\mathbb{R}^m$, the SMACOF algorithm can be summarized as follows:

Initialize embedding coordinate $Z^{(0)}$.
for $k = 1, 2, \cdots$ **do**
Perform multiplicative update

$$Z^{(k)} = \frac{1}{N} B(Z^{(k-1)}) Z^{(k-1)},$$

where $B(Z)$ is an $N \times N$ matrix-valued function with elements

$$b_{ij}(Z) = \begin{cases} \frac{d_X(x_i,x_j)}{\|z_i - z_j\|} & i \neq j \text{ and } \|z_i - z_j\| \neq 0, \\ 0 & i \neq j \text{ and } \|z_i - z_j\| = 0, \\ -\sum_{k\neq i} b_{ik} & i = j. \end{cases}$$

end for

SMACOF iteration is equivalent to a weighted steepest descent with constant step size [32], but due to a special structure of the problem, it guarantees monotonous decrease of the stress function [17]. Other L_p formulations can be solved using iteratively reweighted least-squares (IRLS) techniques [16]. Acceleration of convergence is possible using multiscale and multigrid methods [32] as well as vector extrapolation techniques [96].

Eigenmaps

In the specific case when the shape is equipped with the diffusion distance ($d = d_{X,t}$), the canonical form can be computed by observing the fact that the map $\Phi_{X,t}(x) = (e^{-\lambda_0 t}\phi_0(x), e^{-\lambda_1 t}\phi_1(x), \ldots)$ defined by the eigenvalues and eigenvectors of Laplace–Beltrami operator Δ_X satisfies $d_{X,t}(x, x') = \|\Phi_t(x) - \Phi_t(x')\|_2$.

In other words, $\Phi_{X,t}$ is an *isometric embedding* of $(X, d_{X,t})$ into an infinite-dimensional Euclidean space and can be thought of as an infinite-dimensional canonical form [13, 72]. $\Phi_{X,t}$ is termed *Laplacian eigenmap* [9] or *diffusion map* [40]. Another eigenmap given by $\Psi_X(x) = \left(\lambda_1^{-1/2}\phi_1(x), \lambda_2^{-1/2}\phi_2(x), \dots\right)$, referred to as the *global point signature* (GPS) [98], is an isometric embedding of the commute time metric c_X.

Unlike Elad–Kimmel canonical forms computed by MDS, the eigenmap is uniquely defined (i.e., there are no degrees of freedom related to the isometry in the embedding space) if the Laplace–Beltrami operator has no eigenvalues of multiplicity greater than one. Otherwise, the ambiguity in the definition of the eigenmap is up to switching between the eigenfunction corresponding to the eigenvalues with multiplicity and changes in their signs. More ambiguities arise in cases of symmetric shapes [86]. In general, two eigenmaps may differ by a permutation of coordinates corresponding to simple eigenvalues or by a roto-reflection in the eigensubspaces corresponding to eigenvalues with multiplicities.

For practical comparison of eigenmaps, a finite number k of eigenvectors is used, $\tilde{\Phi}_{X,t} = (e^{-\lambda_0 t}\phi_0, \dots, e^{-\lambda_k t}\phi_k)$. The Euclidean distance on the eigenmap $\tilde{\Phi}_{X,t}$ is thus a numerical approximation to the diffusion metric $d_{X,t}$ using k eigenfunctions of the Laplace–Beltrami operator (similarly, $\tilde{\Psi}_X$ approximates the commute time). For small k, eigenmaps can be compared using ICP. The problem of coordinate permutations must be addressed if eigenvalues of multiplicity greater than one are present. Such an approach is impractical for $k >> 1$.

As an alternative, Rustamov [98] proposed using shape distributions to compare eigenmaps. This method overcomes the aforementioned problem, but lacks the metric properties of a true isometry-invariant metric.

Gromov–Hausdorff Distance

The source of inaccuracy of Elad–Kimmel canonical forms is that it is generally impossible to select a common metric space $(\mathbb{Z}, d_{\mathbb{Z}})$ in which the geometry of any shape can be accurately represented. However, for given two shapes X and Y, the space $(\mathbb{Z}, d_{\mathbb{Z}})$ can be selected in such a way that (X, d) and (Y, δ) can be isometrically embedded into it, the simplest example being the disjoint union $\mathbb{Z} = X \sqcup Y$ of X and Y, with the metric $d_{\mathbb{Z}}|_{X\times X} = d$ and $d_{\mathbb{Z}}|_{Y\times Y} = \delta$. $d_{\mathbb{Z}}|_{X\times Y}$ is defined to minimize the Hausdorff distance between X and Y in $(\mathbb{Z}, d_{\mathbb{Z}})$, resulting in a distance,

$$d_{\mathrm{GH}}(X, Y) = \inf_{d_{\mathbb{Z}}} d_{\mathrm{H}}^{\mathbb{Z}}(X, Y), \tag{41}$$

called the *Gromov–Hausdorff distance*. The Gromov–Hausdorff distance was first proposed by Gromov [55] as a distance between metric spaces and a generalization of the Hausdorff distance and brought into shape recognition by Mémoli and Sapiro [78].

The Gromov–Hausdorff distance satisfies axioms (Ii)–(Iii) with $c_1 = c_2 = 2$ and $b_1 = b_2 = 0$, such that $d_{\mathrm{GH}}(X, Y) = 0$ if and only if X and Y are isometric. More generally, if $d_{\mathrm{GH}}(X, Y) \leq \epsilon$, then X and Y are 2ϵ-isometric, and, conversely, if X and Y are ϵ-isometric, then $d_{\mathrm{GH}}(X, Y) \leq 2\epsilon$ [34].

The Gromov–Hausdorff distance is a generic distance between metric spaces and, in particular, can be used to measure similarity between subsets of the Euclidean metric space, $(X, d_{\mathbb{E}}|_{X\times X})$ and $(Y, d_{\mathbb{E}}|_{Y\times Y})$. In [75], Mémoli showed that the Gromov–Hausdorff distance in the Euclidean space is equivalent to the ICP distance,

$$c \cdot d_{\mathrm{H}}^{\mathbb{E}/\mathrm{Iso}(\mathbb{E})}(X, Y) \leq d_{\mathrm{GH}}((X, d_{\mathbb{E}}|_{X\times X}), (Y, d_{\mathbb{E}}|_{Y\times Y})) \leq d_{\mathrm{H}}^{\mathbb{E}/\mathrm{Iso}(\mathbb{E})}(X, Y),$$

in the sense of equivalence of metrics ($c > 0$ is a constant). (Metric equivalence should not be confused with equality: for example, L_1 and L_2 metrics are equivalent but not equal.) Consequently, (1) if $d_{\mathrm{H}}^{\mathbb{E}/\mathrm{Iso}(\mathbb{E})}(X, Y) \leq \epsilon$, then $(X, d_{\mathbb{E}}|_{X\times X})$ and $(Y, d_{\mathbb{E}}|_{Y\times Y})$ are 2ϵ-isometric; and (2) if $(X, d_{\mathbb{E}}|_{X\times X})$ and $(Y, d_{\mathbb{E}}|_{Y\times Y})$ are ϵ-isometric, then $d_{\mathrm{H}}^{\mathbb{E}/\mathrm{Iso}(\mathbb{E})}(X, Y) \leq c\sqrt{\epsilon}$.

Using the Gromov–Hausdorff distance to compare shapes equipped with diffusion metric allows to benefit from the advantage of the diffusion metric over geodesic one, such as lesser sensitivity to topological noise [28].

Generalized Multidimensional Scaling

For compact metric spaces, the Gromov–Hausdorff distance can also be expressed as

$$d_{\mathrm{GH}}(X, Y) = \frac{1}{2} \inf_{C} \mathrm{dis}\,(C), \tag{42}$$

where the infimum is taken over all correspondence C and $\mathrm{dis}\,(C)$. The two expressions (41) and (42) are equivalent [34].

The advantage of this formulation is that it allows to reduce the computation of the Gromov–Hausdorff distance to finding a minimum-distortion embedding, similarly to the computation of canonical forms by means of MDS. In the discrete setting, given two triangular meshes $\hat{X}$ and $\hat{Y}$ representing the shapes X, Y, let us fix two sufficiently dense finite samplings $P = \{p_1, \ldots, p_m\}$ and $Q = \{q_1, \ldots, q_n\}$ of $\hat{X}$ and $\hat{Y}$, respectively. A discrete correspondence between the shapes is defined as $C = (P \times Q') \cup (Q \times P')$, where $P' = \{p'_1, \ldots, p'_n\}$ and $Q' = \{q'_1, \ldots, q'_m\}$ are some (different) sets of samples on $\hat{X}$ and $\hat{Y}$ corresponding to Q and P, respectively. One can represent C as the union of the graphs of two discrete functions $\varphi : P \to \hat{Y}$ and $\psi : Q \to \hat{X}$, parametrizing the class of all discrete correspondences.

Given two sets P and P' on $\hat{X}$, we can construct an $m \times n$ distance matrix $D(P, P')$, whose elements are the distances $\hat{d}_{\hat{X}}\left(p_i, p'_j\right)$ (e.g., geodesic or diffusion). In these terms, the distortion of the correspondence C can be written as

$$\operatorname{dis}(C) = \left\| \begin{pmatrix} D(P,P) & D(P,P') \\ D(P,P')^{\mathrm{T}} & D(P',P') \end{pmatrix} - \begin{pmatrix} D(Q',Q') & D(Q',Q) \\ D(Q',Q)^{\mathrm{T}} & D(Q,Q) \end{pmatrix} \right\|,$$

where $\|\cdot\|$ is some norm on the space of $(m+n)\times(m+n)$ matrices. The selection of the infinity norm $\|D\|_\infty = \max_{i,j}|d_{ij}|$ is consistent with the Gromov–Hausdorff distance; however, in practice more robust norms like the Frobenius norm $\|D\|_{\mathrm{F}}^2 = \operatorname{trace}(DD^{\mathrm{T}})$ are often preferable (see [23, 74, 78] for discussions on the regularization of the infinity norm in the Gromov–Hausdorff framework by other l_p norms).

The discretization of $\operatorname{dis}(C)$ leads directly to a discretized approximation of the Gromov–Hausdorff distance between shapes, which can be expressed as

$$\hat{d}_{\mathrm{GH}}(\hat{X},\hat{Y}) := \frac{1}{2}\min_{P',Q'} \operatorname{dis}(C).$$

Note that only P' and Q' participate as continuous minimization variables, while P and Q are constants (given samples on the respective shapes). The above minimization problem is solved using a numerical procedure resembling MDS, first introduced in [23, 24] under the name *generalized MDS* (GMDS).

We use barycentric coordinates to represent points on $\hat{X}$ and $\hat{Y}$. In these coordinates, a point p_i lying in a triangle t_i on $\hat{X}$ is represented as a convex combination of the triangle vertices (corresponding to the indices t_i^1, t_i^2, and t_i^3) with the weights $u_i = \left(u_i^1, u_i^2, u_i^3\right)^{\mathrm{T}}$. We will denote by $T = (t_1,\ldots,t_m)^{\mathrm{T}}$ the vector of triangle indices and by $U = (u_1,\ldots,u_m)$ the $3\times m$ matrix of coordinates corresponding to the sampling P. Similarly, the samplings P', Q, and Q' are represented as (T',U'), (S,V), and (S',V'). For the sake of notation simplicity, we are going to use these interchangeably.

It was shown in [26] that a first-order approximation of a geodesic distance between p_i' and p_j' on $\hat{X}$ can be expressed as the quadratic form

$$D_{ij}(P',P') \approx u_i'^{\mathrm{T}} \begin{pmatrix} D_{t_i^1,t_j^1}(P,P) & D_{t_i^1,t_j^2}(P,P) & D_{t_i^1,t_j^3}(P,P) \\ D_{t_i^2,t_j^1}(P,P) & D_{t_i^2,t_j^2}(P,P) & D_{t_i^2,t_j^3}(P,P) \\ D_{t_i^3,t_j^1}(P,P) & D_{t_i^3,t_j^2}(P,P) & D_{t_i^3,t_j^3}(P,P) \end{pmatrix} u_j'.$$

Other distance terms are expressed similarly. Using tensor notation, we can write

$$\operatorname{dis}(C) \approx \left\| (U,U')\mathcal{D}_{\hat{X}}(T,T')(U,U') - (V,V')\mathcal{D}_{\hat{Y}}(S,S')(V,V') \right\|_{\mathrm{F}}^2,$$

where $\mathcal{D}_{\hat{X}}(T,T')$ is a rank four tensor whose ij-th elements are defined as the 3×3 distance matrices above, and $\mathcal{D}_{\hat{Y}}(S,S')$ is defined in a similar way.

The resulting objective function $\operatorname{dis}(C)$ is a fourth-order polynomial with respect to the continuous coordinates U', V', also depending on the discrete index variables T' and S'. However, when all indices and all

coordinate vectors except one, say, u'_i, are fixed, the function becomes convex and quadratic with respect to u'_i. A closed-form minimizer of dis $\left(u'_i\right)$ is found under the constraints $u'_i \geq 0$ and $u'^1_i + u'^2_i + u'^3_i = 1$, guaranteeing that the point p'_i remains within the triangle t'_i. The GMDS minimization algorithm proceeds iteratively by selecting u'_i or v'_i corresponding to the largest gradient of the objective function, updating it according to the closed-form minimizer, and updating the corresponding triangle index to a neighboring one in case the solution is found on the boundary of the triangle. The reader is referred to [26] for further implementation details.

Graph-Based Methods

The minimum-distortion correspondence problem can be formulated as a *binary labeling* problem with uniqueness constraints [110] in a graph with vertices defined as pairs of points and edges defined as quadruplets. Let $\mathcal{V} = \{(x, y) : x \in X, y \in Y\} = X \times Y$ be the set of pairs of points from X and Y, and let $\mathcal{E} = \{((x, y), (x', y')) \in \mathcal{V} \times \mathcal{V} \text{ and } (x, y) \neq (x', y')\}$. A correspondence $C \subset X \times Y$ can be represented as binary labeling $u \in \{0, 1\}^{\mathcal{V}}$ of the graph $(\mathcal{V}, \mathcal{E})$, as follows: $u_{x,y} = 1$ iff $(x, y) \in C$ and 0 otherwise. When using L_2 distortions, the correspondence problem can be reformulated as

$$\min_{u \in \{0,1\}^{\mathcal{V}}} \sum_{((x,y),(x',y')) \in \mathcal{E}} u_{x,y} u_{x',y'} |d_X(x, x') - d_Y(y, y')|^2$$
$$\text{s.t.} \sum_y u_{x,y} \leq 1 \ \ \forall x \in X; \quad \sum_x u_{x,y} \leq 1 \ \ \forall y \in Y. \tag{43}$$

In general, optimization of this energy is NP-hard [53]. One possible approximation of (43) is by relaxing the labels to be in $[0, 1]$. This formulation leads to a non-convex quadratic program with linear constraints [46, 74]. Alternatively, instead of minimizing directly the energy (43), it is possible to maximize a lower bound on it by solving the dual to the linear programming (LP) relaxation of (43), a technique known as *dual decomposition* [110]. This approaches demonstrate good global convergence behavior [65].

Probabilistic Gromov–Hausdorff Distance

The Gromov–Hausdorff framework can be extended to a setting in which pairwise distances are replaced by *distributions* of distances, modeling the intra-class variability shapes (e.g., the fact that different humans have legs of different length) [115]. The pairwise metric difference terms in the correspondence distortion are replaced by probabilities, and the problem is posed as likelihood maximization.

Gromov–Wasserstein Distances

The same way as the Gromov–Hausdorff extends the Hausdorff distance by taking a minimum over all possible metric spaces, $d_{\mathrm{GH}} = \min_{d_{\mathbb{Z}}} d_{\mathrm{H}}^{\mathbb{Z}}$, an extension for the Wasserstein distance of the form

$$\begin{aligned} d_{\mathrm{GW},p}(X,Y) &= \min_{d_{\mathbb{Z}}} d_{\mathrm{W},p}^{\mathbb{Z}}(X,Y) \\ &= \min_{d_{\mathbb{Z}}} \min_{\mu\in\mathcal{M}(\mu_X,\mu_Y)} \left(\int_{X\times Y} d_{\mathbb{Z}}^p(x,y)d\mu(x,y)\right)^{\frac{1}{p}}, \end{aligned} \tag{44}$$

referred to as *Gromov–Wasserstein distance*, was proposed by Mémoli [74]. Here, it is assumed that X and Y are metric measure spaces with metrics d_X, d_Y and measures μ_X, μ_Y. The analogy between the Gromov–Hausdorff and the Gromov–Wasserstein distances is very close: the Hausdorff distance is a distance between subsets of a metric measure space, and the Gromov–Hausdorff distance is a distance between metric spaces. The Wasserstein distance is a distance between subsets of a metric space, and the Gromov–Wasserstein distance is a distance between metric measure spaces.

Numerical Computation

In [74], Mémoli showed that (44) can be alternatively formulated as

$$\begin{aligned} &d_{\mathrm{GW},p}(X,Y) = \\ &\min_{\mu\in\mathcal{M}(\mu_X,\mu_Y)} \left(\int_{X\times Y}\int_{X\times Y} |d_X(x,x') - d_Y(y,y')|^p d\mu(x,y)d\mu(x',y')\right)^{\frac{1}{p}}. \end{aligned} \tag{45}$$

This formulation has an advantage in numerical implementation. Given discrete surfaces $\{x_1,\ldots,x_N\}$ and $\{y_1,\ldots,y_M\}$ with discretized metrics $d_X(x_i,x_{i'})$, $d_Y(y_j,y_{j'})$ and measures $\mu_X(x_i), \mu_Y(y_j)$ (for $i,i' = 1,\ldots,N$ and $j = 1,\ldots,M$), problem (45) can be posed as an optimization problem with NM variables and $N+M$ linear constraints:

$$\begin{aligned} \min_{\mu} & \sum_{i,i'=1}^{N}\sum_{j,j'=1}^{M} \mu_{ij}\mu_{i'j'}|d_X(x_i,x_{i'}) - d_Y(y_j,y_{j'})|^p \\ \text{s.t.} \quad & \mu_{ij} \in [0,1] \\ & \sum_{i=1}^{N} \mu_{ij} = \mu_Y(y_j) \\ & \sum_{j=1}^{M} \mu_{ij} = \mu_X(x_i). \end{aligned}$$

Shape DNA

Reuter et al. [95] proposed using the Laplace–Beltrami *spectrum* (i.e., eigenvalues $\lambda_0, \lambda_1, \ldots$ of Δ_X) as shape descriptors, referred to as *shape DNA*. Laplace–Beltrami spectrum is isometry invariant; however, there may exist shapes which are *isospectral* (have equal eigenvalues) but non-isometric. This fact was first conjectured by Kac [62] and shown by example in [54]. Thus, the equivalence class of isospectral shapes to which the shape DNA approach is invariant is wider than the class of isometries. The exact relations between these classes are currently unknown.

6 Partial Similarity

In many situations, it happens that while two objects are not similar, some of their parts are. Such a situation is common, for example, in the face recognition application, where the quality of facial images (or surfaces in the case of 3D face recognition) can be degraded by acquisition imperfections, occlusions, and the presence of facial hair. Semantically, we can say that two objects are partially similar if they have significant similar parts. If one is able to detect such parts, the degree of partial similarity can be evaluated.

We define a part of a shape (X, d_X) simply as its subset $X' \subset X$ equipped with the restricted metric $d_X|_{X' \times X'}$. According to this definition, every part of a shape is also a shape. We will denote by $\Sigma(X) \subset 2^X$ the set of all admissible parts, satisfying (1) $\Sigma(X)$ is nonempty; (2) $\Sigma(X)$ is closed under complement, i.e., if $X' \in \Sigma(X)$, then $X \setminus X' \in \Sigma(X)$; and (3) $\Sigma(X)$ is closed under countable unions, i.e., any countable union of parts from $\Sigma(X)$ is also an admissible part in $\Sigma(X)$. Formally, the set of all parts of X is a *σ-algebra*. An equivalent representation of a part is by means of a binary indicator function, $p : X \to \{0, 1\}$, assuming the value of one for each $x \in X'$ and zero otherwise. We will see the utility of such a definition in the sequel.

Significance

The *significance* of a part is a function on $\Sigma(X)$ assigning each part a number quantifying its "importance." We denote significance by σ and demand that (1) σ is nonnegative; (2) $\sigma(\emptyset) = 0$; and (3) σ is countably additive, i.e., $\sigma\left(\bigcup_i X'_i\right) = \sum_i \sigma\left(X'_i\right)$ for every countable union of parts in $\Sigma(X)$. Formally, significance is a finite *measure* on X. As in the case of similarity, the notion of significance is application dependent. The most straightforward way to define significance is by identifying it with the *area*

$$\sigma(X') = \int_{X'} da$$

or the *normalized area*

$$\sigma(X') = \frac{\int_{X'} da}{\int_X da}.$$

of the part. However, such a definition might deem equally important a large flat region and a region rich in features if they have the same area, while it is clear that the latter one would usually be more informative. A better approach is to interpret significance as the amount of information about the entire shape contained in its part, quantified, e.g., as the ability to discriminate the shape from a given corpus of other shapes given only its part. Such a definition leads to a weighted area measure, where the weighting reflects the *discriminativity density* of each point and is constructed similarly to the term frequency-inverse document frequency (TF-IDF) weighting commonly used in text retrieval [2].

Regularity

Another quantity characterizing the importance of a part is its *regularity*, which we model as a scalar function $\rho : \Sigma(X) \to \mathbb{R}$ [18, 19]. In general, we would like the part to be simple, i.e., if two parts contain the same amount of information (are equally significant), we would prefer the simpler one, following Ockham's *pluralitas non est ponenda sine necessitate* principle. What is exactly meant by "regular" and "simple" is again application dependent. In many applications, an acceptable definition of regularity is the deviation of a shape from some perfectly regular one. For example, in image processing and computer vision, regularity is commonly expressed using the *shape factor*

$$\rho(X') = \frac{4\pi \int_{X'} da}{\left(\int_{\partial X'} ds\right)^2},$$

or the ratio between the area of X' and the squared length of its boundary. Because of the isoperimetric inequality in the plane, this ratio is always less or equal to one, with the equality achieved by a circle, which is arguably a very regular shape. Shape factor can be readily extended to non-Euclidean shapes, where, however, there is no straightforward analogy of the isoperimetric inequality. Consequently, two equally regular shapes might have completely different topology, e.g., one might have numerous disconnected components, while the other having only one (Fig. 8).

A remedy can be in regarding regularity as a purely topological property, counting, for example, the number of disconnected components of a part. Topological

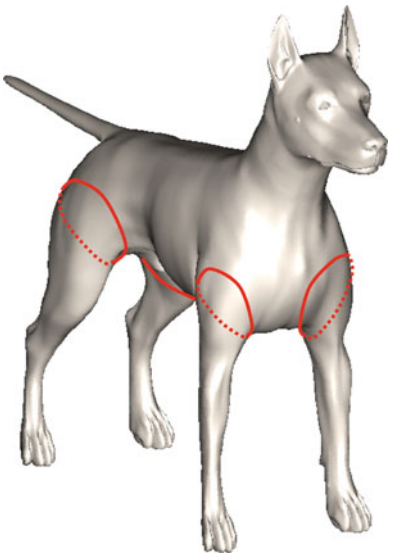


Fig. 8 Large shape factor does not necessarily imply regularity in non-Euclidean shapes. Here, the upper body of the dog and the four legs have the same area and the same boundary length (*red contours*) and, hence, the same shape factor. However, the upper body is arguably more regular than the four disconnected legs (Reproduced from [26])

regularity can be expressed in terms of the *Euler characteristic*, which using the Gauss–Bonnet identity becomes

$$\rho(X') = 2\pi\chi(X') = \int_{X'} K da + \int_{\partial X'} k_g ds,$$

where K is the Gaussian curvature of X and k_g is the geodesic curvature of $\partial X'$.

Partial Similarity Criterion

In this terminology, the problem of partial similarity of two shapes X and Y can be thought finding two parts $X' \in \Sigma(X)$ and $Y' \in \Sigma(Y)$ simultaneously maximizing regularity, significance, and similarity. Since a part of a shape is also a shape, the latter can be quantified using any shape similarity (Since we use *dissimilarity*, we will maximize $-d(X', Y')$.) criterion appropriate for the application, e.g., the Gromov–Hausdorff distance. This can be written as the following multi-criterion optimization problem [18, 19, 27]

$$\max_{\substack{X' \in \Sigma(X) \\ Y' \in \Sigma(Y)}} (\rho(X'), \rho(Y'), \sigma(X'), \sigma(Y'), -d(X', Y')),$$

where maximum is understood as a point in the criterion space, such that no other point has all the criteria larger simultaneously. Such a maximum is said to be *Pareto efficient* and is not unique. The solution of this multi-criterion maximization problem can be interpreted as a *set-valued* partial similarity criterion. Since such

criteria are not mutually comparable, the problem should be converted into a scalar maximization problem

$$\max_{\substack{X' \in \Sigma(X) \\ Y' \in \Sigma(Y)}} \lambda_r(\rho(X') + \rho(Y')) + \lambda_s(\sigma(X') + \sigma(Y')) - d(X', Y'), \tag{46}$$

where λ_r and λ_s are positive scalars reflecting the tradeoff between regularity, significance, and dissimilarity.

Computational Considerations

Direct solution of problem (46) involves searching over the space of all parts of X and Y, which has combinatorial complexity. However, the problem can be relaxed to maximization in continuous variables if binary parts are allowed to be *fuzzy*. Formally, a fuzzy part is obtained by letting the binary indicator functions assume values on the interval $[0, 1]$. Such functions are called *membership functions* in the fuzzy set theory terminology. The optimization problem becomes [28]

$$\max_{\substack{p: X \to [0,1] \\ q: Y \to [0,1]}} \lambda_r(\rho(p) + \rho(q)) + \lambda_s(\sigma(p) + \sigma(q)) - d(p, q),$$

where $\rho(p), \sigma(p)$ and $d(p, q)$ are the fuzzy counterparts of the regularity, significance, and dissimilarity terms. The significance of a fuzzy part p is simply

$$\sigma(p) = \int_X p \, d\sigma.$$

The regularity term is somewhat more involved as it involves integration along the part boundary, which does not exist in case of a fuzzy part. However, the following relaxation is available [35]

$$\rho(p) = \frac{4\pi \int_X p \, da}{\left(\int_X \|\nabla p\| \delta\left(p - \frac{1}{2}\right) da \right)^2},$$

with δ being the Dirac delta function. This fuzzy version of the shape factor converges to the original definition when p approaches a binary indicator function. The dissimilarity term needs to be modified to involve the membership function. The most straightforward way to do so is by defining a weighted dissimilarity between the entire shapes X and Y with p and q serving as the weights. For example, using $p(x)da(x)$ and $q(y)da(y)$ as the respective measures on X and Y, the Wasserstein distance incorporates the weights in a natural way.

7 Self-Similarity and Symmetry

An important particular case of shape similarity is the similarity of shape with itself, which is commonly referred to as *symmetry*. The latter notion is intimately related with that of invariance.

Rigid Symmetry

Computation of exact and approximate symmetries has been extensively studied in the Euclidean sense [3,6,83,118]. A shape X is said to be symmetric if there exists a nontrivial Euclidean isometry $f \in \mathrm{Iso}(\mathbb{R}^3)$ to which it is invariant, i.e., $f(X) = X$. Such an isometry is called a symmetry of X. True symmetries, like true isometries, are a mere idealization not existing in practice. In real applications, we might still find approximate symmetries. The degree of asymmetry of a Euclidean isometry f can be quantified as a distance between X and $f(X)$ in $\mathbb{R}^3$, e.g.,

$$\mathrm{asym}(f) = d_{\mathrm{H}}^{\mathbb{R}^3}(X, f(X)).$$

Intrinsic Symmetry

A symmetry f restricted to X defines a *self-isometry* of X, i.e., $f|_X \in \mathrm{Iso}(X)$. Therefore, an alternative definition of an approximate symmetry could be an ϵ-isometry, with the distortion quantifying the degree of asymmetry. Such a definition requires approximate symmetries to be automorphisms of X, yet its main advantage is the fact that it can be extended beyond the Euclidean case (Fig. 9). In fact, identifying the symmetry group with the isometry group $\mathrm{Iso}(X, d_X)$ of the shape X with some intrinsic (e.g., geodesic or diffusion) metric d_X, a nonrigid equivalent of symmetries is defined, while setting $d_X = d_{\mathrm{E}}|_{X\times X}$ the standard Euclidean symmetries are obtained [92]. Approximate symmetries with respect to any metric can be computed as local minima of the distortion function in embedding X into itself. Computationally, the process can be carried out using GMDS.

Spectral Symmetry

An alternative to this potentially heavy computation is due to Ovsjanikov et al. [87] and is based on the elegant observation that for any simple (A simple eigenfunction is one corresponding an eigenvalues with multiplicity one.) eigenfunction ϕ_i of the Laplace–Beltrami operator, a reflection symmetry f satisfies $\phi_i \circ f = \pm\phi_i$. This allows parametrize reflection symmetries by *sign sequences* $\mathbf{s} = \{s_1, s_2, \dots\}$, $s_i \in \{\pm 1\}$, such that $\phi_i \circ f = s_i\phi_i$.

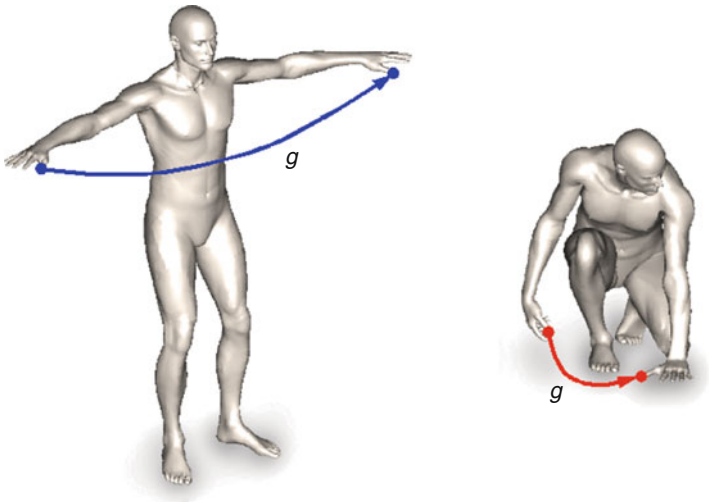


Fig. 9 Symmetry defined as a metric-preserving automorphism (self-isometry) of X allows extending the standard notion of Euclidean symmetry to nonrigid shapes (Reproduced from [92])

Given a sign sequence, the eigenmap $\Phi_{\mathbf{s}}(x) = \left\{s_1\lambda_1^{-1/2}\phi_1(x), s_2\lambda_2^{-1/2}\phi_2(x), \dots\right\}$ is defined. Symmetries are detected by evaluating the asymmetry

$$\text{asym}(\mathbf{s}) = \max_{x\in X} \min_{x'\in X} \|\Phi_{\mathbf{s}}(x') - \Phi(x)\|$$

of different sign sequences and keeping those having asym $\leq \epsilon$. The symmetry itself corresponding to a sequence $\mathbf{s}$ is recovered as

$$f(x) = \arg\min_{x'\in X} \|\Phi_{\mathbf{s}}(x') - \Phi(x)\|,$$

and is an ϵ self-isometry of X in the sense of the commute time metric. While it can be made relatively computationally simple, this method is limited to global reflection symmetries only.

Partial Symmetry

In many cases, a shape does not have symmetries as a whole, yet possesses parts that are symmetric. Adopting the notion of partial similarity defined in Sect. 6, one can think of a part $X' \subset X$ and a *partial symmetry* $f : X' \times X'$ as of a Pareto-efficient tradeoff between asymmetry asym(f), part significance $\sigma(X')$,

and regularity $\rho(X')$. Partial symmetries are found similarly to the computation of partial similarity of two distinct shapes.

Repeating Structure

Another important particular case of self-similarity is repeating *regular structure*. Shapes possessing regular structure can be divided into self-similar patches (*structural elements*) forming some regular patterns, e.g., a grid. State-of-the-art methods [89, 108**?**] can detect structured repetitions in extrinsic geometry if the Euclidean transformations between repeated patches exhibit group-like behavior. In case of nonrigid and deformable shapes, however, the problem is challenging since no apparent structure is visible to simple Euclidean probes in the absence of repetitive Euclidean transformations to describe the shape. A general solution for the detection of intrinsic regular structure is still missing, though particular cases have been recently addressed in [29].

8 Feature-Based Methods

Another class of methods, referred to as *feature based*, uses local information to describe the shape, perform matching, or compute similarity. The popularity of these methods has increased following the success of the scale-invariant feature transform (SIFT) [70] and similar algorithms [8, 71] in image analysis and computer vision application.

Feature Descriptors

In essence, feature-based methods try to represent the shape as a collection of local *feature descriptors*. This is typically done in two steps first, selecting robust and representative points (*feature detection*) and computing the local shape representation at these points (*feature description*).

Feature Detection

One of the main requirements on a feature detector is that the points it selects are (1) *repeatable*, i.e., in two instances of a shape, ideally the same set of corresponding points is detected, and (2) *informative*, i.e., the information contained in these points is sufficient to, e.g., distinguish the shape from others.

In the most trivial case, no feature detection is performed and the feature descriptor is computed at all the points of the shape or at some regularly sampled subset thereof. The descriptor in this case is usually termed *dense*. Dense descriptors bypass the problem of repeatability at the price of increased computational cost and potentially introducing many unimportant points that clutter the shape representation.

Many geometric feature detection paradigms come from the image analysis community, such as finding points with high derivatives (e.g., the *Harris operator* [30, 52, 57]) or local maxima in a scale space (e.g., *difference of Gaussians* (DOG) [119] or local maxima of the heat kernel [49]).

Feature Description

Given a set of feature points (or, in the case of a dense descriptor, all the points on the shape), a local descriptor is then computed. An ideal feature descriptor should be (1) invariant under the class of transformations a shape can undergo and (2) informative. One of the most known feature descriptors is *spin image* [4, 5, 61], describing the neighborhood of a point by fitting an oriented coordinate system at the point. Belongie and Malik introduced the *shape context descriptor* [11], describing the structure of the shape as relations between a point to the rest of the point. Given the coordinates of a point x on the shape, the shape context descriptor is constructed as a histogram of the direction vectors from x to the rest of the point, $y - x$. Typically, a log-polar histogram is used. Because of dependence on the embedding coordinates, such a descriptor is not deformation invariant. Other descriptors exist based on local patches [82], local moments [38] and volume descriptors [50], spherical harmonics [101], and contour and edge structures [64,88]. Zaharescu et al. [119] proposed using as a local descriptor the histogram of gradients of a function (e.g., Gaussian curvature) defined in a neighborhood of a point, similarly to the *histogram of gradients* (HOG) [43] and SIFT [70] techniques used in computer vision.

Because considering local geometry, feature descriptors are usually not very susceptible to nonrigid deformations of the shape. Nevertheless, there exist several geometric descriptors which are invariant to isometric deformations by construction. Examples include descriptors based on histograms of local geodesic distances [27], conformal factors [12], and heat kernels [106], described in the following in more details.

Heat Kernel Signatures

Sun et al. [106] proposed the *heat kernel signature* (HKS), defined as the diagonal of the heat kernel. Given some fixed time values $t_1, \ldots, t_n$, for each point x on the shape, the HKS is an n-dimensional descriptor vector

$$p(x) = (K_{t_1}(x,x), \ldots, K_{t_n}(x,x)). \tag{47}$$

The HKS descriptor is deformation invariant, captures local geometric information at multiple scales, is insensitive to topological noise, is informative (if the Laplace–Beltrami operator of a shape is non-degenerate, then any continuous map that preserves the HKS at every point must be an isometry), and is easily computed across different shape representations solving the eigenproblem described in section "Diffusion Distance."

Scale-Invariant Heat Kernel Signatures

A disadvantage of the HKS is its dependence on the global scale of the shape. If X is globally scaled by β, the corresponding HKS is $\beta^{-2}K_{\beta^{-2}t}(x,x)$. In some cases, it is possible to remove this dependence by *global* normalization of the shape. A *scale-invariant HKS* (SI-HKS) based on *local* normalization was proposed in [33]. By using a logarithmic scale-space $t = \alpha^\tau$, the scaling of X by β results in HKS amplitude scaling and shift by $2\log_\alpha \beta$. This effect is undone by the following sequence of transformations,

$$\begin{aligned} p_{dif}(x) \;=\; & (\log K_{\alpha^{\tau_2}}(x,x) - \log K_{\alpha^{\tau_1}}(x,x), \ldots, \log K_{\alpha^{\tau_m}}(x,x) \\ & - \log K_{\alpha^{\tau_{m-1}}}(x,x)),\ \hat{p}(x) = |(\mathcal{F}p_{dif}(x))(\omega_1,\ldots,\omega_n)|, \end{aligned} \tag{48}$$

where $\mathcal{F}$ is the discrete Fourier transform and $\omega_1,\ldots,\omega_n$ denotes a set of frequencies at which the transformed vector is sampled. Taking differences of logarithms removes the scaling constant, and the Fourier transform converts the scale-space shift into a complex phase, which is removed by taking the absolute value.

Bags of Features

One of the notable advantages of feature-based approaches is the possibility of representing a shape as a collection of primitive elements ("geometric words") and using the well-developed methods from text search such as the *bag of features* (BOF) (or *bag of words*) paradigm [37, 103]. Such approaches are widely used in image retrieval and have been introduced more recently to shape analysis [29, 109]. The bag of features representation is usually compact and easy to store and compare, which makes such approaches suitable for large-scale shape retrieval.

The construction of a bag of features is usually performed in a few steps, depicted in Fig. 10. First, the shape is represented as a collection of local feature descriptors (either dense or computed at a set of stable points following an optional stage of feature detection). Second, the descriptors are represented by *geometric words* from a *geometric vocabulary* using vector quantization. The geometric vocabulary is a set of representative descriptors, precomputed in advance. This way, each descriptor is replaced by the index of the closest geometric word in the vocabulary. Computing the histogram of the frequency of occurrence of geometric words gives the bag of features. Alternatively, a two-dimensional histogram of co-occurrences of pairs of geometric words (*geometric expressions*) can be used [29]. Shape similarity is computed as a distance between the corresponding bags of features.

Combining Global and Local Information

Another use of local descriptors is in combination with global (metric) information, in an extension of the Gromov–Hausdorff framework. Given two shapes X, Y with

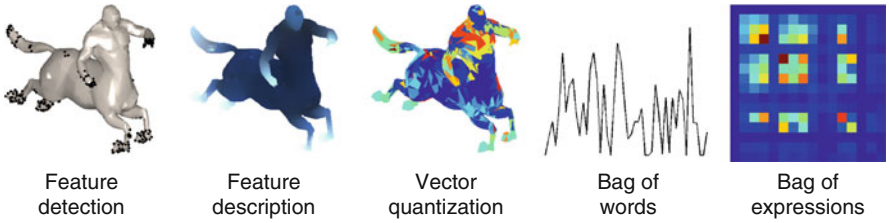


Fig. 10 Feature-based shape analysis algorithm (Reproduced from [29])

metrics d_X, d_Y and descriptors p_X, p_Y, the quality of correspondence $C \subseteq X \times Y$ is measured using global geometric distortion as well as local matching of descriptors,

$$\operatorname{dis}(C) = \sup_{(x,y),(x',y')\in C} |d_X(x,x') - d_Y(y,y')| + \beta \sup_{(x,y)\in C} \|p_X(x) - p_Y(y)\|,$$

where $\beta > 0$ is some parameter. This L_∞ formulation can be replaced by a more robust L_2 version. As the descriptors, texture [105, 107] or geometric information [46, 115] can be used.

The minimum-distortion correspondence can be found by an extension of the GMDS algorithm described in section "Generalized Multidimensional Scaling" [107] or graph labeling [105,110,115] described in section "Graph-Based Methods." The probabilistic extension of the Gromov–Hausdorff distance can be applied to this formulation as well [115].

9 Conclusion

In this chapter, the problem of invariant shape similarity was presented through the prism of metric geometry. It was shown that by representing shapes as metric spaces allows to reduce the similarity problem to isometry-invariant comparison of metric spaces. The particular choice of the metric results in different isometry groups and, hence, different invariance classes. The construction of Euclidean, geodesic, and diffusion metrics was presented, and their theoretical properties were highlighted in Sect. 2. Based on these notions, different shape similarity criteria and distances were presented in Sect. 5, fitting well under the metric umbrella. Computational aspects related to shape and metric discretization were discussed in Sects. 3 and 4, and computation of full and partial similarity was discussed in Sects. 5 and 6. In Sect. 8, feature-based methods were discussed. For further detailed discussion of these and related subjects, the reader is referred to the book [25].

Cross-References

- ▸ Shape Spaces
- ▸ Variational Methods in Shape Analysis
- ▸ Image Segmentation with Shape Priors: Explicit Versus Implicit Representations

References

1. Adams, C.C., Franzosa, R.: Introduction to Topology: Pure and Applied. Prentice-Hall, Harlow (2008)
2. Aizawa, A.: An information-theoretic perspective of TF–IDF measures. Inf. Process. Manag. **39**(1), 45–65 (2003)
3. Alt, H., Mehlhorn, K., Wagener, H., Welzl, E.: Congruence, similarity, and symmetries of geometric objects. Discret. Comput. Geom. **3**, 237–256 (1988)
4. Andreetto, M., Brusco, N., Cortelazzo, G.M.: Automatic 3D modeling of textured cultural heritage objects. Trans. Image Process. **13**(3), 335–369 (2004)
5. Assfalg, J., Bertini, M., Pala, P., Del Bimbo, A.: Content-based retrieval of 3D objects using spin image signatures. Trans. Multimed. **9**(3), 589–599 (2007)
6. Atallah, M.J.: On symmetry detection. IEEE Trans. Comput. **c-34**(7), 663–666 (1985)
7. Aurenhammer, F.: Voronoi diagramsa survey of a fundamental geometric data structure. ACM Comput. Surv. **23**(3), 345–405 (1991)
8. Bay, H., Tuytelaars, T., Van Gool, L.: SURF: speeded up robust features. In: Proceedings of European Conference on Computer Vision (ECCV), Graz, pp. 404–417 (2006)
9. Belkin, M., Niyogi, P.: Laplacian eigenmaps for dimensionality reduction and data representation. Neural Comput. **13**, 1373–1396 (2003). Introduction of Laplacian embeddings
10. Bellman, R.E.: Dynamic Programming. Dover, New York (2003)
11. Belongie, S., Malik, J., Puzicha, J.: Shape matching and object recognition using shape contexts. IEEE Trans. Pattern Anal. Mach. Intell. (PAMI) **24**, 509–522 (2002)
12. Ben-Chen, M., Weber, O., Gotsman, C.: Characterizing shape using conformal factors. In: Proceedings of 3DOR, Crete (2008)
13. Bérard, P., Besson, G., Gallot, S.: Embedding Riemannian manifolds by their heat kernel. Geom. Funct. Anal. **4**(4), 373–398 (1994)
14. Bernstein, M., de Silva, V., Langford, J.C., Tenenbaum, J.B.: Graph approximations to geodesics on embedded manifolds, Technical report (2000)
15. Besl, P.J., McKay, N.D.: A method for registration of 3D shapes. IEEE Trans. Pattern Anal. Mach. Intell. (PAMI) **14**(2), 239–256 (1992). Introduction of ICP
16. Bjorck, A.A.: Numerical Methods for Least Squares Problems. Society for Industrial Mathematics, Philadelphia (1996)
17. Borg, I., Groenen, P.: Modern Multidimensional Scaling – Theory and Applications. Comprehensive Overview of MDS Problems and Their Numerical Solution. Springer, New York (1997)
18. Bronstein, A.M., Bronstein, M.M.: Not only size matters: regularized partial matching of nonrigid shapes. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR Workshops), Anchorage (2008)
19. Bronstein, A.M., Bronstein, M.M.: Regularized partial matching of rigid shapes. In: Proceedings of European Conference on Computer Vision (ECCV), Marseille, pp. 143–154 (2008)
20. Bronstein, A.M., Bronstein, M.M., Kimmel, R.: Expression-invariant 3D face recognition. In: Proceedings of Audio and Video-Based Biometric Person Authentication, Guildford. Lecture Notes in Computer Science, vol. 2688, pp. 62–69 (2003). Springer, Berlin. 3D face recognition using metric model

21. Bronstein, A.M., Bronstein, M.M., Kimmel, R.: On isometric embedding of facial surfaces into S3. In: Proceedings of International Conference Scale Space and PDE Methods in Computer Vision, Hofgeismar. Lecture Notes in Computer Science, vol. 3459, pp. 622–631. Springer, New York (2005). MDS with spherical geometry
22. Bronstein, A.M., Bronstein, M.M., Kimmel, R.: Three-dimensional face recognition. Int. J. Comput. Vis. (IJCV) **64**(1), 5–30 (2005). 3D face recognition using metric model
23. Bronstein, A.M., Bronstein, M.M., Kimmel, R.: Efficient computation of isometry-invariant distances between surfaces. SIAM J. Sci. Comput. **28**(5), 1812–1836 (2006). Computation of the Gromov-Hausdorff distance using GMDS
24. Bronstein, A.M., Bronstein, M.M., Kimmel, R.: Generalized multidimensional scaling: a framework for isometry-invariant partial surface matching. Proc. Natl. Acad. Sci. (PNAS) **103**(5), 1168–1172 (2006). Introduction of generalized MDS
25. Bronstein, A.M., Bronstein, M.M., Kimmel, R.: Robust expression-invariant face recognition from partially missing data. In: Proceedings of European Conference on Computer Vision (ECCV), Graz, pp. 396–408 (2006). 3D face recognition with partially missing data
26. Bronstein, A.M., Bronstein, M.M., Kimmel, R.: Numerical Geometry of Non-rigid Shapes. Springer, New York (2008). First systematic treatment of non-rigid shapes
27. Bronstein, A.M., Bronstein, M.M., Bruckstein, A.M., Kimmel, R.: Partial similarity of objects, or how to compare a centaur to a horse. Int. J. Comput. Vis. (IJCV) **84**(2), 163–183 (2009)
28. Bronstein, A.M., Bronstein, M.M., Kimmel, R., Mahmoudi, M., Sapiro, G.: A Gromov-Hausdorff framework with diffusion geometry for topologically-robust non-rigid shape matching. Int. J. Comput. Vis. (IJCV) **89**(2–3), 266–286 (2010)
29. Bronstein, A.M., Bronstein, M.M., Ovsjanikov, M., Guibas, L.J.: Shape Google: a computer vision approach to invariant shape retrieval. In: Proceedings of Non-rigid Shapes and Deformable Image Alignment (NORDIA) (2009)
30. Bronstein, A.M., Bronstein, M.M., Bustos, B., Castellani, U., Crisani, M., Falcidieno, B., Guibas, L.J., Isipiran, I., Kokkinos, I., Murino, V., Ovsjanikov, M., Patané, G., Spagnuolo, M., Sun, J.: Robust feature detection and description benchmark. In: Proceedings of 3DOR, Firenze (2010)
31. Bronstein, M.M., Bronstein, A.M.: Shape recognition with spectral Distances. IEEE Trans. Pattern Anal. Mach. Intell. (PAMI) (2010, in press)
32. Bronstein, M.M., Bronstein, A.M., Kimmel, R., Yavneh, I.: Multigrid multidimensional scaling. Numer. Linear Algebra Appl. **13**(2–3), 149–171 (2006). Multigrid solver for MDS problems
33. Bronstein, M.M., Kokkinos, I.: Scale-invariant heat kernel signatures for non-rigid shape recognition. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco (2010)
34. Burago, D., Burago, Y., Ivanov, S.: A Course in Metric Geometry. Graduate Studies in Mathematics, vol. 33. AMS, Providence (2001). Systematic introduction to metric geometry
35. Chan, T.F., Vese, L.A.: A level set algorithm for minimizing the Mumford-Shah functional in image processing. In: IEEE Workshop on Variational and Level Set Methods, Beijing, pp. 161–168 (2001)
36. Chen, Y., Medioni, G.: Object modeling by registration of multiple range images. In: Proceedings of Conference on Robotics and Automation, Sacramento (1991). Introduction of ICP
37. Chum, O., Philbin, J., Sivic, J., Isard, M., Zisserman, A.: Total recall: automatic query expansion with a generative feature model for object retrieval. In: Proceedings of International Conference on Computer Vision (ICCV), Rio de Janeiro (2007)
38. Clarenz, U., Rumpf, M., Telea, A.: Robust feature detection and local classification for surfaces based on moment analysis. Trans. Vis. Comput. Graph. **10**(5), 516–524 (2004)
39. Coifman, R.R., Lafon, S.: Diffusion maps. Appl. Comput. Harmon. Anal. **21**(1), 5–30 (2006). Definition of diffusion distance

40. Coifman, R.R., Lafon, S., Lee, A.B., Maggioni, M., Nadler, B., Warner, F., Zucker, S.W.: Geometric diffusions as a tool for harmonic analysis and structure definition of data: diffusion maps. Proc. Natl. Acad. Sci. (PNAS) **102**(21), 7426–7431 (2005). Introduction of diffusion maps and diffusion distances
41. Cox, T.F., Cox, M.A.: Multidimensional Scaling. Chapman & Hall, London (1994)
42. Crandal, M.G., Lions, P.-L.: Viscosity solutions of Hamilton–Jacobi equations. Trans. Am. Math. Soc. **277**, 1–43 (1983)
43. Dalai, N., Triggs, B.: Histograms of oriented gradients for human Detection. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), San Diego (2005)
44. De Leeuw, J.: Applications of convex analysis to multidimensional scaling. In: Recent Developments in Statistics, pp. 133–145. North-Holland, Amsterdam (1977)
45. Du, Q., Faber, V., Gunzburger, M.: Centroidal Voronoi tessellations: applications and algorithms. SIAM Rev. **41**(4), 637–676 (2006)
46. Dubrovina, A., Kimmel, R.: Matching shapes by eigendecomposition of the Laplace-Beltrami operator. In: Proceedings of 3DPVT, Paris (2010)
47. Elad, A., Kimmel, R.: Bending invariant representations for surfaces. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Kauai, pp. 168–174 (2001). Introduction of canonical forms
48. Elad, A., Kimmel, R.: On bending invariant signatures for surfaces. IEEE Trans. Pattern Anal. Mach. Intell. (PAMI) **25**(10), 1285–1295 (2003). Introduction of canonical forms
49. Gebal, K., Bærentzen, J.A., Aanæs, H., Larsen, R.: Shape analysis using the auto diffusion function. Comput. Graph. Forum **28**(5), 1405–1413 (2009)
50. Gelfand, N., Mitra, N.J., Guibas, L.J., Pottmann, H.: Robust global registration. In: Proceedings of Symposium on Geometry Processing (SGP), Vienna (2005)
51. Gersho, A., Gray, R.M.: Vector Quantization and Signal Compression. Kluwer, Boston (1992)
52. Glomb, P.: Detection of interest points on 3D data: extending the Harris operator. In: Computer Recognition Systems 3. Advances in Soft Computing, vol. 57, pp. 103–111. Springer, Berlin/Heidelberg (2009)
53. Gold, S., Rangarajan, A.: A graduated assignment algorithm for graph matching. IEEE Trans. Pattern Anal. Mach. Intell. (PAMI) **18**, 377–388 (1996)
54. Gordon, C., Webb, D.L., Wolpert, S.: One cannot hear the shape of the drum. Bull. AMS **27**(1), 134–138 (1992). Example of isospectral but non-isometric shapes
55. Gromov, M.: Structures Métriques Pour les Variétés Riemanniennes. In: Textes Mathématiques, vol. 1 (1981). Introduction of the Gromov-Hausdorff distance
56. Gu, X., Gortler, S., Hoppe, H.: Geometry images. In: Proceedings of SIGGRAPH, San Antonio, pp. 355–361 (2002)
57. Harris, C., Stephens, M.: A combined corner and edge detection. In: Proceedings of Fourth Alvey Vision Conference, Manchester, pp. 147–151 (1988)
58. Hausdorff, F.: Grundzüge der Mengenlehre, Definition of the Hausdorff Distance. Verlag Veit & Co, Leipzig (1914)
59. Hochbaum, D.S., Shmoys, D.B.: A best possible heuristic for the k-center problem. Math. Oper. Res. **10**, 180–184 (1985)
60. Indyk, P., Thaper, N.: Fast image retrieval via embeddings. In: 3rd International Workshop on Statistical and Computational Theories of Vision, Nice (2003)
61. Johnson, A.E., Hebert, M.: Using spin images for efficient object recognition in cluttered 3D scenes. IEEE Trans. Pattern Anal. Mach. Intell. (PAMI) **21**(5), 433–449 (1999)
62. Kac, M.: Can one hear the shape of a drum? Am. Math. Mon. **73**, 1–23 (1966). Kac's conjecture about isospectral but non-isometric shapes
63. Kimmel, R., Sethian, J.A.: Computing geodesic paths on manifolds. Proc. Natl. Acad. Sci. (PNAS) **95**(15), 8431–8435 (1998)
64. Kolomenkin, M., Shimshoni, I., Tal, A.: On edge detection on surfaces. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Miami (2009)

65. Komodakis, N., Paragios, N., Tziritas, G.: MRF optimization via dual decomposition: message-passing revisited. In: Proceedings of International Conference on Computer Vision (ICCV), Rio de Janeiro (2007)
66. Leibon, G., Letscher, D.: Delaunay triangulations and Voronoi diagrams for Riemannian manifolds. In: Proceedings of Symposium on Computational Geometry, Hong Kong, pp. 341–349 (2000)
67. Lévy, B.: Laplace-Beltrami eigenfunctions towards an algorithm that "understands" geometry. In: International Conference on Shape Modeling and Applications, Matsushima (2006). The use of Laplace-Belrami operator for shape analysis and synthesis
68. Lloyd, S.P.: Least squares quantization in PCM. Bell telephone laboratories paper (1957)
69. Losasso, F., Hoppe, H., Schaefer, S., Warren, J.: Smooth geometry images. In: Proceedings of Symposium on Geometry Processing (SGP), Aachen, pp. 138–145 (2003)
70. Lowe, D.: Distinctive image features from scale-invariant keypoint. Int. J. Comput. Vis. (IJCV) **60**, 91–110 (2004)
71. Matas, J., Chum, O., Urban, M., Pajdla, T.: Robust wide-baseline stereo from maximally stable extremal regions. Image Vis. Comput. **22**(10), 761–767 (2004)
72. Mateus, D., Horaud, R.P., Knossow, D., Cuzzolin, F., Boyer, E.: Articulated shape matching using Laplacian eigenfunctions and unsupervised point registration. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Anchorage (2008)
73. Max, J.: Quantizing for minimum distortion. IEEE Trans. Inf. Theory **6**(1), 7–12 (1960)
74. Mémoli, F.: On the use of Gromov-Hausdorff distances for shape Comparison. In: Proceedings of Point Based Graphics, Prague (2007). Definition of the Gromov-Wasserstein distance
75. Mémoli, F.: Gromov-Hausdorff distances in Euclidean spaces. In: Proceedings of Non-rigid Shapes and Deformable Image Alignment (NORDIA) (2008). Relation of Gromov-Hausdorff distances in Euclidean spaces to Hausdorff and ICP distances
76. Mémoli, F., Sapiro, G.: Fast computation of weighted distance functions and geodesics on implicit hyper-surfaces. J. Comput. Phys. **173**(1), 764–795 (2001)
77. Mémoli, F., Sapiro, G.: Distance functions and geodesics on submanifolds of Rd and point clouds. SIAM J. Appl. Math. **65**(4), 1227 (2005)
78. Mémoli, F., Sapiro, G.: A theoretical and computational framework for isometry invariant recognition of point cloud data. Found. Comput. Math. **5**, 313–346 (2005). First use of the Gromov-Hausdorff distance in shape recognition
79. Meyer, M., Desbrun, M., Schroder, P., Barr, A.H.: Discrete differential-geometry operators for triangulated 2-manifolds. In: Visualization and Mathematics III, pp. 35–57 (2003). Cotangent weights discretization of the Laplace-Beltrami operator
80. Mitra, N.J., Bronstein, A.M., Bronstein, M.M.: Intrinsic regularity detection in 3D geometry. In: Proceedings of European Conference on Computer Vision (ECCV), Heraklion (2010)
81. Mitra, N.J., Gelfand, N., Pottmann, H., Guibas, L.: Registration of point cloud data from a geometric optimization perspective. In: Proceedings of Eurographics Symposium on Geometry Processing, Aachen, pp. 23–32 (2004). Analysis of ICP algorithms from optimization standpoint
82. Mitra, N.J., Guibas, L.J., Giesen, J., Pauly, M.: Probabilistic fingerprints for shapes. In: Proceedings of Symposium on Geometry Processing (SGP), Cagliari, Sardinia (2006)
83. Mitra, N.J., Guibas, L.J., Pauly, M.: Partial and approximate symmetry detection for 3D geometry. ACM Trans. Graph. (TOG) **25**(3), 560–568 (2006)
84. Nash, J.: The imbedding problem for Riemannian manifolds. Ann. Math. **63**, 20–63 (1956). Nash embedding theorem
85. Osada, R., Funkhouser, T., Chazelle, B., Dobkin, D.: Shape distributions. ACM Trans. Graph. (TOG) **21**(4), 807–832 (2002). Introduction of the shape distributions method for rigid shapes
86. Ovsjanikov, M., Sun, J., Guibas, L.: Global intrinsic symmetries of Shapes. Comput. Graph. Forum **27**, 1341–1348 (2008). Spectral method for non-rigid symmetry detection
87. Ovsjanikov, M., Sun, J., Guibas, L.J.: Global intrinsic symmetries of shapes. In: Proceedings of Symposium on Geometry Processing (SGP), Copenhagen, pp. 1341–1348 (2008)

88. Pauly, M., Keiser, R., Gross, M.: Multi-scale feature extraction on point-sampled surfaces. Comput. Graph. Forum **22**, 281–289 (2003)
89. Pauly, M., Mitra, N.J., Wallner, J., Pottmann, H., Guibas, L.J.: Discovering structural regularity in 3D geometry. ACM Trans. Graph. (TOG) **27**(3), 43 (2008)
90. Peyre, G., Cohen, L.: Surface segmentation using geodesic centroidal Tesselation. In: Proceedings of International Symposium on 3D Data Processing Visualization Transmission, Thessaloniki, pp. 995–1002 (2004)
91. Pinkall, U., Polthier, K.: Computing discrete minimal surfaces and their conjugates. Exp. Math. **2**(1), 15–36 (1993). Cotangent weights discretization of the Laplace-Beltrami operator
92. Raviv, D., Bronstein, A.M., Bronstein, M.M., Kimmel, R.: Symmetries of non-rigid shapes. In: Proceedings of Workshop on Non-rigid Registration and Tracking Through Learning (NRTL) (2007)
93. Raviv, D., Bronstein, A.M., Bronstein, M.M., Kimmel, R.: Full and partial symmetries of non-rigid shapes. Int. J. Comput. Vis. (IJCV) **89**(1), 18–39 (2010)
94. Reuter, M., Biasotti, S., Giorgi, D., Patanè, G., Spagnuolo, M.: Discrete Laplace-Beltrami operators for shape analysis and segmentation. Comput. Graph. **33**, 381–390 (2009). FEM approximation of the Laplace-Beltrami operator
95. Reuter, M., Wolter, F.-E., Peinecke, N.: Laplace-Beltrami spectra as "shape-DNA" of surfaces and solids. Comput. Aided Des. **38**(4), 342–366 (2006). Shape recognition using Laplace-Beltrami spectrum
96. Rosman, G., Bronstein, A.M., Bronstein, M.M., Sidi, A., Kimmel, R.: Fast multidimensional scaling using vector extrapolation. Technical report CIS-2008-01, Department of Computer Science, Technion, Israel (2008). Introduction of vector extrapolation methods for MDS problems
97. Rubner, Y., Guibas, L.J., Tomasi, C.: The earth movers distance, multi-dimensional scaling, and color-based image retrieval. In: Proceedings of the ARPA Image Understanding Workshop, New Orleans, pp. 661–668 (1997)
98. Rustamov, R.M.: Laplace-Beltrami eigenfunctions for deformation invariant shape representation. In: Proceedings of Symposium on Geometry Processing (SGP), Barcelona, pp. 225–233 (2007). Introduction of GPS embedding
99. Sander, P., Wood, Z., Gortler, S., Snyder, J., Hoppe, H.: Multichart geometry images. In: Proceedings of Symposium on Geometry Processing (SGP), Aachen, pp. 146–155 (2003)
100. Sethian, J.A.: A fast marching level set method for monotonically advancing fronts. Proc. Natl. Acad. Sci. (PNAS) **93**(4), 1591–1595 (1996)
101. Shilane, P., Funkhauser, T.: Selecting distinctive 3D shape descriptors for similarity retrieval. In: Proceedings of Shape Modelling and Applications, Matsushima (2006)
102. Shirdhonkar, S., Jacobs, D.W.: Approximate earth movers distance in linear time. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Anchorage (2008)
103. Sivic, J., Zisserman, A.: Video Google: a text retrieval approach to object matching in videos. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Nice (2003)
104. Spira, A., Kimmel, R.: An efficient solution to the eikonal equation on parametric manifolds. Interfaces Free Bound. **6**(4), 315–327 (2004)
105. Starck, J., Hilton, A.: Correspondence labelling for widetimeframe free-form surface matching. In: Proceedings of International Conference on Computer Vision (ICCV), Rio de Janeiro (2007)
106. Sun, J., Ovsjanikov, M., Guibas, L.J.: A concise and provably informative multi-scale signature based on heat diffusion. In: Proceedings of Symposium on Geometry Processing (SGP), Berlin (2009)
107. Thorstensen, N., Keriven, R.: Non-rigid shape matching using geometry and photometry. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Miami (2009)

108. Thrun, S., Wegbreit, B.: Shape from symmetry. In: Proceedings of International Conference on Computer Vision (ICCV), Beijing (2005)
109. Toldo, R., Castellani, U., Fusiello, A.: Visual vocabulary signature for 3D object retrieval and partial matching. In: Proceedings of 3DOR, Munich (2009)
110. Torresani, L., Kolmogorov, V., Rother, C.: Feature correspondence via graph matching: models and global optimization. In: Proceedings of European Conference on Computer Vision (ECCV), Marseille, pp. 596–609 (2008)
111. Tsai, Y.R., Cheng, L.T., Osher, S., Zhao, H.K.: Fast sweeping algorithms for a class of Hamilton-Jacobi equations. SIAM J. Numer. Anal. (SINUM) **41**(2), 673–694 (2003)
112. Tsitsiklis, J.N.: Efficient algorithms for globally optimal trajectories. IEEE Trans. Autom. Control **40**(9), 1528–1538 (1995)
113. Tutte, W.T.: How to draw a graph. Proc. Lond. Math. Soc. **13**(3), 743–768 (1963). Tutte Laplacian operator
114. Walter, J., Ritter, H.: On interactive visualization of high-dimensional data using the hyperbolic plane. In: Proceedings of International Conference on Knowledge Discovery and Data Mining (KDD), Edmonton, pp. 123–131 (2002). MDS with hyperbolic geometry
115. Wang, C., Bronstein, M.M., Paragios, N.: Discrete minimum distortion correspondence problems for non-rigid shape matching, Research report 7333, INRIA (2010)
116. Wardetzky, M., Mathur, S., Kälberer, F., Grinspun, E.: Discrete Laplace operators: no free lunch. In: Conference on Computer Graphics and Interactive Techniques (2008). Analysis of different discretizations of the Laplace-Beltrami operator
117. Weber, O., Devir, Y.S., Bronstein, A.M., Bronstein, M.M., Kimmel, R.: Parallel algorithms for approximation of distance maps on parametric surfaces. ACM Trans. Graph. (TOG) **27**(4), 1–16 (2008)
118. Wolter, J.D., Woo, T.C., Volz, R.A.: Optimal algorithms for symmetry detection in two and three dimensions. Vis. Comput. **1**, 37–48 (1985)
119. Zaharescu, A., Boyer, E., Varanasi, K., Horaud, R.: Surface feature detection and description with applications to mesh matching. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Miami (2009)
120. Zhang, H.: Discrete combinatorial Laplacian operators for digital geometry processing. In: SIAM Conference on Geometric Design, pp. 575–592 (2004). Combinatorial Laplace-Beltrami operator
121. Zhao, H.K.: Fast sweeping method for eikonal equations. Math. Comput. **74**, 603–627 (2005)

Image Segmentation with Shape Priors: Explicit Versus Implicit Representations

Daniel Cremers

Contents

1 Introduction 1910
Image Analysis and Prior Knowledge 1910
Explicit Versus Implicit Shape Representation 1911
2 Image Segmentation via Bayesian Inference 1913
3 Statistical Shape Priors for Parametric Shape Representations 1915
Linear Gaussian Shape Priors 1916
Nonlinear Statistical Shape Priors 1918
4 Statistical Priors for Level Set Representations 1921
Shape Distances for Level Sets 1922
Invariance by Intrinsic Alignment 1923
Kernel Density Estimation in the Level Set Domain 1926
Gradient Descent Evolution for the Kernel Density Estimator 1929
Nonlinear Shape Priors for Tracking a Walking Person 1930
5 Dynamical Shape Priors for Implicit Shapes 1931
Capturing the Temporal Evolution of Shape 1931
Level Set-Based Tracking via Bayesian Inference 1932
Linear Dynamical Models for Implicit Shapes 1934
Variational Segmentation with Dynamical Shape Priors 1935
6 Parametric Representations Revisited: Combinatorial Solutions for Segmentation with Shape Priors 1937
7 Conclusion 1938
Cross-References 1940
References 1940

D. Cremers (✉)
Department of Computer Science, Technische Universität München, Garching, Germany

© Springer Science+Business Media New York 2015

O. Scherzer (ed.), *Handbook of Mathematical Methods in Imaging*,
DOI 10.1007/978-1-4939-0790-8_40

1 Introduction

Image Analysis and Prior Knowledge

Image segmentation is among the most studied problems in image understanding and computer vision. The goal of image segmentation is to partition the image plane into a set of meaningful regions. Here *meaningful* typically refers to a semantic partitioning where the computed regions correspond to individual objects in the observed scene. Unfortunately, generic purely low-level segmentation algorithms often do not provide the desired segmentation results, because the traditional low-level assumptions like intensity or texture homogeneity and strong edge contrast are not sufficient to separate objects in a scene.

To overcome these limitations, researchers have proposed to impose prior knowledge into low-level segmentation methods. In the following, we will review methods which allow to impose knowledge about the *shape* of objects of interest into segmentation processes.

In the literature there exist various definitions of the term *shape*, from the very broad notion of shape of Kendall [54] and Bookstein [5] where shape is whatever remains of an object when similarity transformations are factored out (i.e., a geometrically normalized version of a gray value image) to more specific notions of shape referring to the geometric outline of an object in 2D or 3D. In this work, we will adopt the latter view and refer to an object's silhouette or boundary as its shape. Intentionally we will leave the exact mathematical definition until later, as different representations of geometry actually imply different definitions of the term *shape*.

One can distinguish various kinds of shape knowledge:

- Low-level shape priors which typically simply favor shorter boundary length, i.e., curves with shorter boundary have lower shape energy, where boundary length can be measured in various ways [4, 6, 49, 53, 69].
- Mid-level shape priors which favor, for example, thin and elongated structures, thereby facilitating the segmentation of roads in satellite imagery or of blood vessels in medical imagery [44, 70, 78].
- High-level shape priors which favor similarity to previously observed shapes, such as hand shapes [22,36,50], silhouettes of humans [26,29], or medical organs like the heart, the prostate, the lungs, or the cerebellum [58, 82, 84, 99].

There exists a wealth of works on shape priors for image segmentation. It is beyond the scope of this article to provide a complete overview of existing work. Instead, we will present a range of representative works – with many of the examples taken from the author's own work – and discuss their advantages and shortcomings. Some of these works are formulated in a probabilistic setting where the challenge is to infer the most likely shape given an image and a set of training shapes. Typically the segmentation is formulated as an optimization problem.

One can distinguish two important challenges:

1. The modeling challenge: How do we formalize distances between shapes? What probability distributions do we impose? What energies should we minimize?
2. The algorithmic challenge: How do we minimize the arising cost function? Are the computed solutions globally optimal? If they are not globally optimal, how sensitive are solutions with respect to the initialization?

Explicit Versus Implicit Shape Representation

A central question in the modeling of shape similarity is that of how to represent a shape. Typically one can distinguish between *explicit* and *implicit* representations. In the former case, the boundary of the shape is represented explicitly – in a spatially continuous setting, this could be a polygon or a spline curve. In a spatially discrete setting this could be a set of edgles (edge elements) forming a regular grid. Alternatively, shapes can be represented implicitly in the sense that one labels all points in space as being part of the interior or the exterior of the object. In the spatially continuous setting, the optimization of such implicit shape representations is solved by means of partial differential equations. Among the most popular representatives are the level set method [39, 72] or alternative convex relaxation techniques [11]. In the spatially discrete setting, implicit representations have become popular through the graph cut methods [7, 49]. More recently, researchers have also advocated hybrid representations where objects are represented both explicitly and implicitly [90]. Table 1 provides an overview of a few representative works on image segmentation based on explicit and implicit representations of shape.

Figure 1 shows examples of shape representations using an explicit parametric representation by spline curves (spline control points are marked as black boxes), implicit representations by a signed distance function or a binary indicator function, and an explicit discrete representation (4th image).

As we shall see in the following, the choice of shape representation has important consequences on the class of objects that can be modeled, the type of energy that can be minimized, and the optimality guarantees that can be obtained. Among the goals of this article is to put in contrast various shape representations and discuss their advantages and limitations. In general one observes that:

Table 1 Shapes can be represented explicitly or implicitly, in a spatially continuous or a spatially discrete setting. More recently, researchers have adopted hybrid representations [90], where objects are represented both in terms of their interior (implicitly) and in terms of their boundary (explicitly)

	Spatially continuous	Spatially discrete	
Explicit	Polygons [22, 102], splines [3, 36, 53]	Edgel labeling and dyn. progr. [1, 74, 80, 87, 89]	Hybrid repres. and LP relaxation [90]
Implicit	Level set methods [39, 72], convex relaxation [11, 31]	Graph cut methods [6, 49]	

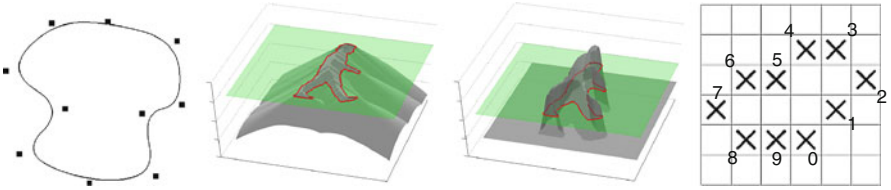


Fig. 1 Examples of shape representations by means of a parametric spline curve (1st image), a signed distance function (2nd image), a binary indicator function (3rd image), and an explicit discrete representation (4th image)

- Implicit representations are easily generalized to shapes in arbitrary dimension. Respective algorithms (level set methods, graph cuts, convex relaxation techniques) straightforwardly extend to three or more dimensions. Instead, the extension of explicit shape representations to higher dimensions is by no means straightforward: The notion of arc-length parameterization of curves does not extend to surfaces. Moreover, the discrete polynomial-time shortest-path algorithms [1,85,89] which allow to optimally identify pairwise correspondence of points on either shape do not directly extend to minimal-surface algorithms.
- Implicit representations are easily generalized to arbitrary shape topology. Since the implicit representation merely relies on a labeling of space (as being inside or outside the object), the topology of the shape is not constrained. Both level set and graph cut algorithms can therefore easily handle objects of arbitrary topology. Instead, for spatially continuous parametric curves, modeling the transition from a single closed curve to a multiply connected object boundary requires sophisticated splitting and merging techniques [38,60,61,65]. Similarly, discrete polynomial-time algorithms are typically constrained to finding open [1,20,23] or closed curves [86,89].
- Explicit boundary representations allow to capture the notion of *point correspondence* [47,85,89]. The correspondence between points on either of two shapes and the underlying correspondence of semantic parts is of central importance to human notions of shape similarity. The determination of optimal point correspondences, however, is an important combinatorial challenge, especially in higher dimensions.
- For explicit representations, the modeling of shape similarity is often more straightforward and intuitive. For example, for two shapes parameterized as spline curves, the linear interpolation of these shapes also gives rise to a spline curve and often captures the human intuition of an *intermediate shape*. Instead, the linear interpolation of implicit representations is generally not straightforward: Convex combinations of binary-valued functions are no longer binary-valued. And convex combinations of signed distance functions are generally no longer a signed distance function. Figure 2 shows examples of a linear interpolations of spline curves and a linear interpolations of signed distance functions. Note that the linear interpolation of signed distance functions may give rise to intermediate silhouettes of varying topology.

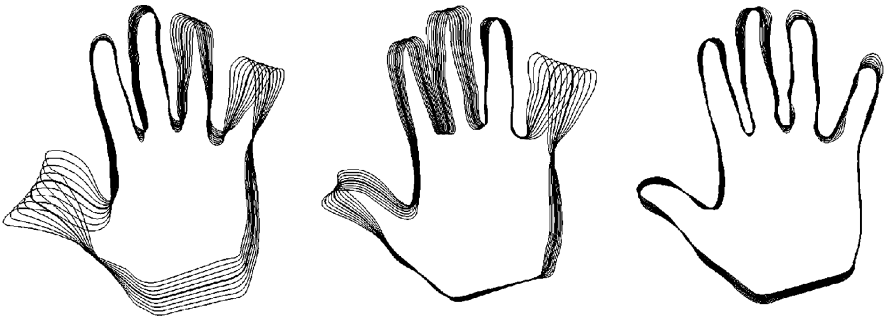


Fig. 2 The linear interpolation of spline-based curves (shown here along the first three eigenmodes of a shape distribution) gives rise to a families of intermediate shapes

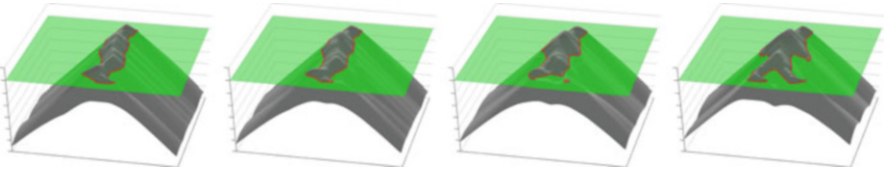


Fig. 3 This figure shows the linear interpolation of the signed distance functions associated with two human silhouettes. The interpolation gives rise to intermediate shapes and allows changes of the shape topology. Yet, the linear combination of two signed distance functions is generally no longer a signed distance function

In the following, we will give an overview over some of the developments in the domain of shape priors for image segmentation. In Sect. 2, we will review a formulation of image segmentation by means of Bayesian inference which allows the fusion of input data and shape knowledge in a single energy minimization framework (Fig. 3). In Sect. 3, we will discuss a framework to impose statistical shape priors in a spatially continuous parametric representation. In Sect. 4, we discuss methods to impose statistical shape priors in level set based image segmentation. In Sect. 5, we discuss statistical models which allow to represent the temporal evolution of shapes and can serve as dynamical priors for image sequence segmentation. And lastly, in Sect. 6, we will present recent developments to impose elastic shape priors in a manner which allows to compute globally optimal shape-consistent segmentations in polynomial time.

2 Image Segmentation via Bayesian Inference

Over the last decades Bayesian inference has become an established paradigm to tackle data analysis problems – see [30, 105] for example. Given an input image $I : \Omega \to \mathbb{R}$ on a domain $\Omega \subset \mathbb{R}^2$, a segmentation $\mathcal{C}$ of the image plane Ω can be computed by maximizing the posterior probability:

$$\mathcal{P}(\mathcal{C} \,|\, I) = \frac{\mathcal{P}(I \,|\, \mathcal{C})\, \mathcal{P}(\mathcal{C})}{\mathcal{P}(I)}, \tag{1}$$

where $\mathcal{P}(I \,|\, \mathcal{C})$ denotes the data likelihood for a given segmentation $\mathcal{C}$ and $\mathcal{P}(\mathcal{C})$ denotes the prior probability which allows to impose knowledge about which segmentations are a priori more or less likely.

Maximizing the posterior distribution can be performed equivalently by minimizing the negative logarithm of (1) which gives rise to an energy or cost function of the form

$$E(\mathcal{C}) = E_{\text{data}}(\mathcal{C}) \; + \; E_{\text{shape}}(\mathcal{C}), \tag{2}$$

where $E_{\text{data}}(\mathcal{C}) = -\log \mathcal{P}(I \,|\, \mathcal{C})$ and $E_{\text{shape}}(\mathcal{C}) = -\log \mathcal{P}(\mathcal{C})$ are typically referred to as *data fidelity term* and *regularizer* or *shape prior*. By maximizing the posterior, one aims at computing the most likely solution given data and prior. Of course there exist alternative strategies of either computing solutions corresponding to the mean of the distribution rather than its mode or of retaining the entire posterior distribution in order to propagate multiple hypotheses over time, as done, for example, in the context of particle filtering [3].

Over the years various data terms have been proposed. In the following, we will simply use a piecewise-constant approximation of the input intensity I [69]:

$$E_{\text{data}}(\mathcal{C}) = \sum_{i=1}^{k} \int_{\Omega_i} \big(I(x) - \mu_i\big)^2 dx, \tag{3}$$

where the regions $\Omega_1, \ldots, \Omega_k$ are pairwise disjoint regions separated by the boundary $\mathcal{C}$ and μ_i denotes the average of I over the region Ω_i:

$$\mu_i = \frac{1}{|\Omega_i|} \int_{\Omega_i} I(x)\, dx. \tag{4}$$

More sophisticated data terms based on color likelihoods [8, 57, 103] or texture likelihoods [2, 30] are conceivable.

A glance into the literature indicates that the most prominent image segmentation methods rely on a rather simple geometric shape prior E_{shape} which energetically favors shapes with shorter boundary length [4, 53, 69], a penalizer which – in a spatially discrete setting – dates back at least as far as the Ising model for ferromagnetism [52]. There are several reasons for the popularity of length constraints in image segmentation. Firstly, solid objects in our world indeed tend to be spatially compact. Secondly, such length constraints are mathematically well studied. They give rise to well-behaved models and algorithms – mean curvature motion in a continuous setting and low-order Markov random fields and submodular cost functions in the discrete setting.

Nevertheless, the preference for a shorter boundary is clearly a very simplistic shape prior. In many applications the user may have a more specific knowledge about what kinds of shapes are likely to arise in a given segmentation task. For example, in biology one may want to segment cells that all have a rather specific size and shape. In medical imaging one may want to segment organs that all have a rather unique shape – up to a certain variability – and preserve a specific spatial relationship with respect to other organs. In satellite imagery one may be most interested in segmenting thin and elongated roads, or in the analysis of traffic scenes from a driving vehicle, the predominant objects may be cars and pedestrians. In the following sections, we will discuss ways to impose such *higher-level* shape knowledge into image segmentation methods.

3 Statistical Shape Priors for Parametric Shape Representations

Among the most straightforward ways to represent a shape is to model its outline as a parametric curve. An example is a simple closed spline curve $\mathcal{C} \in \mathcal{C}^k(\mathbb{S}^1, \Omega)$ of the form

$$\mathcal{C}(s) = \sum_{i=1}^{n} p_i \, B_i(s), \tag{5}$$

where $p_i \in \mathbb{R}^2$ denote a set of spline control points and B_i a set of spline basis functions of degree k [19, 36, 43, 66]. In the special case of linear basis functions, we simply have a polygonal shape, used, for example, in [102]. With increasing number of control points, we obtain a more and more detailed shape representation – see Fig. 4. It shows one of the nice properties of parametric shape representations: The representation is quite *compact* in the sense that very detailed silhouettes can be represented by a few real-valued variables.

Given a family of m shapes, each represented by a spline curve of a fixed number of n control points, we can think of these training shapes as a set $\{z_1, \ldots, z_m\}$ of control point vectors:

Fig. 4 Spline representation of a hand shape (*left*) with increasing resolution

$$z_i = (p_{i1}, \ldots, p_{in}) \in \mathbb{R}^{2n}, \tag{6}$$

where we assume that all control point vectors are normalized with respect to translation, rotation, and scaling [41].

With this contour representation, the image segmentation problem boils down to computing an optimal spline control point vector $z \in \mathbb{R}^{2n}$ for a given image. The segmentation process can be constrained to familiar shapes by imposing a statistical shape prior computed from the set of training shapes.

Linear Gaussian Shape Priors

Among the most popular shape prior is based on the assumption that the training shapes are Gaussian distributed – see for example [22, 36, 55]. There are several reasons for the popularity of Gaussian distributions. Firstly, according to the central limit theorem, the average of a large number of i.i.d. random variables is approximately Gaussian distributed – so if the observed variations of shape were created by independent processes, then one could expect the overall distribution to be approximately Gaussian. Secondly, the Gaussian distribution can be seen as a second-order approximation of the true distribution. And thirdly, the Gaussian distribution gives rise to a convex quadratic cost function that allows for easy minimization.

In practice, the number of training shapes m is often much smaller than the number of dimensions $2n$. Therefore, the estimated covariance matrix Σ is degenerate with many zero eigenvalues and thus not invertible. As introduced in [36], a regularized covariance matrix is given by

$$\Sigma_\perp = \Sigma + \lambda_\perp \left(I - V V^t\right), \tag{7}$$

where V is the matrix of eigenvectors of Σ. In this way, we replace all zero eigenvalues of the sample covariance matrix Σ by a constant $\lambda_\perp \in [0, \lambda_r]$, where λ_r denotes the smallest nonzero eigenvalue of Σ. (Note that the inverse $\Sigma_\perp^{-1}$ of the regularized covariance matrix defined in (7) fundamentally differs from the pseudoinverse, the former scaling components in degenerate directions by $\lambda_\perp^{-1}$ while the latter scaling them by 0.) In [68] it was shown that $\lambda_\perp$ can be computed from the true covariance matrix by minimizing the Kullback–Leibler divergence between the exact and the approximated distribution. Yet, since we do not have the exact covariance matrix but merely a *sample* covariance matrix, the reasoning for determining $\lambda_\perp$ suggested in [68] is not justified.

The Gaussian shape prior is then given by

$$\mathcal{P}(z) = \frac{1}{|2\pi\Sigma_\perp|^{1/2}} \exp\left(-\frac{1}{2}(z-\bar{z})^t \, \Sigma_\perp^{-1} \, (z-\bar{z})\right), \tag{8}$$

where $\bar{z}$ denotes the mean control point vector.

Based on the Gaussian shape prior, we can define a shape energy that is invariant to similarity transformations (translation, rotation, and scaling) by

$$E_{\text{shape}}(z) = -\log \mathcal{P}(\hat{z}), \tag{9}$$

where $\hat{z}$ is the shape vector upon similarity alignment with respect to the training shapes:

$$\hat{z} = \frac{R(z - z_0)}{|R(z - z_0)|}, \tag{10}$$

where the optimal translation z_0 and rotation R can be written as functions of z [36]. As a consequence, we can minimize the overall energy

$$E(z) = E_{\text{data}}(C(z)) + E_{\text{shape}}(z) \tag{11}$$

using gradient descent in z. For details on the numerical minimization, we refer to [25, 36].

Figure 5 shows several intermediate steps in a gradient descent evolution on the energy (2) combining the piecewise constant intensity model (3) with a Gaussian shape prior constructed from a set of sample hand shapes. Note how the similarity-

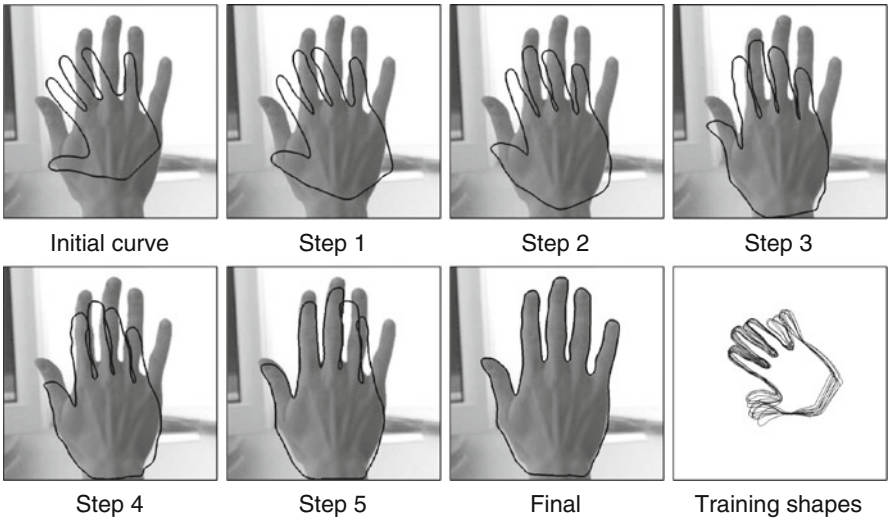


Fig. 5 Evolution of a parametric spline curve during gradient descent on the energy (2) combining the piecewise constant intensity model (3) with a Gaussian shape prior constructed from a set of sample hand shapes (*lower right*). Note that the shape prior is by construction invariant to similarity transformations. As a consequence, the contour easily undergoes translation, rotation, and scaling as these do not affect the energy

Fig. 6 Gradient descent evolution of a parametric curve from initial to final with similarity-invariant shape prior. The statistical shape prior permits a reconstruction of the hand silhouette in places where it is occluded

invariant shape prior (9) constrains the evolving contour to hand-like shapes without constraining its translation, rotation, or scaling.

Figure 6 shows the gradient descent evolution with the same shape prior for an input image of a partially occluded hand. Here the missing part of the silhouette is recovered through the statistical shape prior. These evolutions demonstrate that the curve converges to the correct segmentation over rather large spatial distance, an aspect which is characteristic for region-based cost functions like (3).

Nonlinear Statistical Shape Priors

The shape prior (9) was based on the assumption that the training shapes are Gaussian distributed. For collections of real-world shapes, this is generally not the case. For example, the various silhouettes of a rigid 3D object obviously form a three-dimensional manifold (given that there are only three degrees of freedom in the observation process). Similarly, the various silhouettes of a walking person essentially correspond to a one-dimensional manifold (up to small fluctuations). Furthermore, the manifold of shapes representing deformable objects like human persons are typically very low-dimensional, given that the observed 3D structure only has a small number of joints.

Rather than learning the underlying low-dimensional representation (using principal surfaces or other manifold learning techniques), we can simply estimate arbitrary shape distributions by reverting to nonlinear density estimators – *nonlinear* in the sense that the permissible shapes are not simply given by a weighted sum of eigenmodes. Classical approaches for estimating nonlinear distributions are the Gaussian mixture model or the Parzen–Rosenblatt kernel density estimator – see Sect. 4.

An alternative technique is to adapt recent kernel learning methods to the problem of density estimation [28]. To this end, we approximate the training shapes by a Gaussian distribution, not in the input space but rather upon transformation $\psi : \mathbb{R}^{2n} \rightarrow Y$ to some generally higher-dimensional *feature space* Y:

$$\mathcal{P}_{\psi}(z) \propto \exp\left(-\frac{1}{2}\left(\psi(z)-\psi_0\right)^t \Sigma_{\psi}^{-1}\left(\psi(z)-\psi_0\right)\right). \tag{12}$$

As before, we can define the corresponding shape energy as

$$E(z) = -\log \mathcal{P}_\psi(\hat{z}), \tag{13}$$

with $\hat{z}$ being the similarity-normalized shape given in (10). Here ψ_0 and Σ_ψ denote the mean and covariance matrix computed for the transformed shapes:

$$\psi_0 = \frac{1}{m}\sum_{i=1}^{m}\psi(z_i), \quad \Sigma_\psi = \frac{1}{m}\sum_{i=1}^{m}(\psi(z_i)-\psi_0)(\psi(z_i)-\psi_0)^\top, \tag{14}$$

where Σ_ψ is again regularized as in (7).

As shown in [28], the energy $E(z)$ in (13) can be evaluated without explicitly specifying the nonlinear transformation ψ. It suffices to define the corresponding Mercer kernel [24, 67]:

$$k(x,y) := \langle \psi(x), \psi(y) \rangle, \quad \forall x, y \in \mathbb{R}^{2n}, \tag{15}$$

representing the scalar product of pairs of transformed points $\psi(x)$ and $\psi(y)$. In the following, we simply chose a Gaussian kernel function of width σ:

$$k(x,y) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}}\exp\left(-\frac{||x-y||^2}{2\sigma^2}\right). \tag{16}$$

It was shown in [28] that the resulting energy can be seen as a generalization of the classical Parzen–Rosenblatt estimators. In particular, the Gaussian distribution in feature space Y is fundamentally different from the previously presented Gaussian distribution in the input space $\mathbb{R}^{2n}$. Figure 7 shows the level lines of constant shape energy computed from a set of left- and right-hand silhouettes, displayed

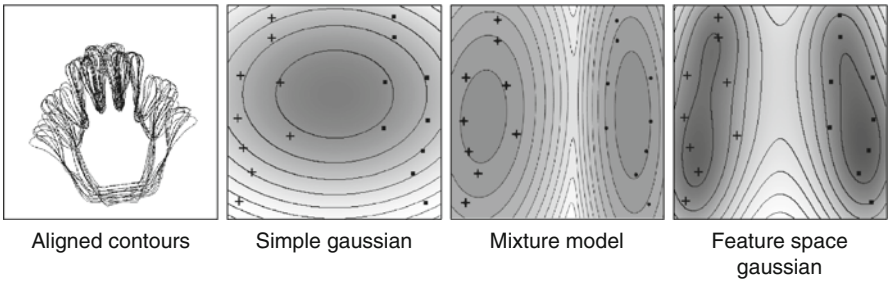


Fig. 7 Model comparison. Density estimates for a set of left (•) and right (+) hands, projected onto the first two principal components. From *left to right*: aligned contours, simple Gaussian, mixture of Gaussians, and Gaussian in feature space (13). In contrast to the mixture model, the Gaussian in feature space does not require an iterative (sometimes suboptimal) fitting procedure

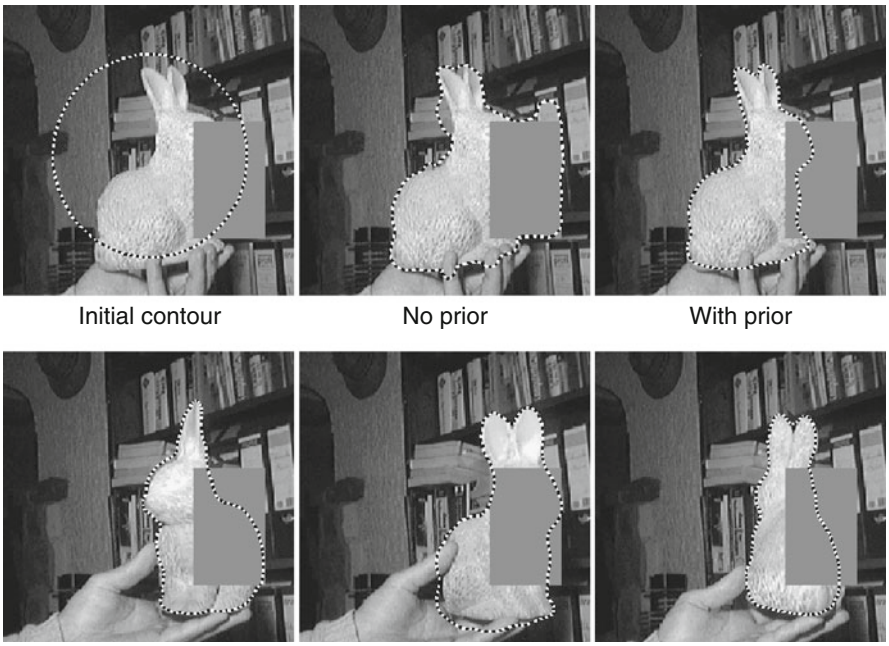


Fig. 8 Tracking a familiar object over a long image sequence with a nonlinear statistical shape prior. A single shape prior constructed from a set of sample silhouettes allows the emergence of a multitude of familiar shapes, permitting the segmentation process to cope with background clutter and partial occlusions

in a projection onto the first two eigenmodes of the distribution. While the linear Gaussian model gives rise to elliptical level lines, the Gaussian mixture and the nonlinear Gaussian allow for more general non-elliptical level lines. In contrast to the mixture model, however, the nonlinear Gaussian does not require an iterative parameter estimation process, nor does it require or assume a specific number of Gaussians.

Figure 8 shows screenshots of contours computed for an image sequence by gradient descent on the energy (11) with the nonlinear shape energy (13) computed from a set of 100 training silhouettes. Throughout the entire sequence, the object of interest was occluded by an artificially introduced rectangle. Again, the shape prior allows to cope with spurious background clutter and to restore the missing parts of the object's silhouette. Two-dimensional projections of the training data and evolving contour onto the first principal components, shown in Fig. 9, demonstrate how the nonlinear shape energy constrains the evolving shape to remain close to the training shapes.

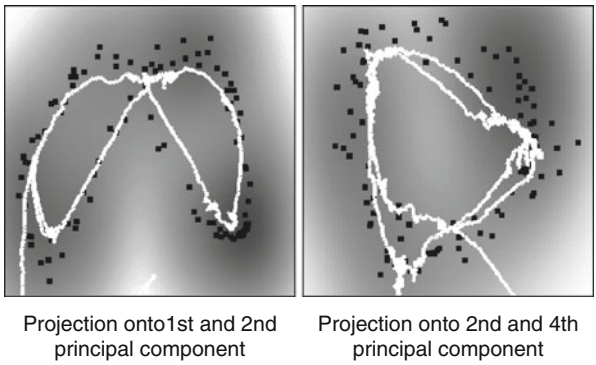


Fig. 9 Tracking sequence from Fig. 8 visualized. Training data (•), estimated energy density (*shaded*), and the contour evolution (*white curve*) in appropriate 2D projections. The evolving contour – see Fig. 8 – is constrained to the domains of low energy induced by the training data

4 Statistical Priors for Level Set Representations

Parametric representations of shape as those presented above have numerous favorable properties; in particular, they allow to represent rather complex shapes with a few parameters, resulting in low memory requirements and low computation time. Nevertheless, the explicit representation of shape has several drawbacks:

- The representation of explicit shapes typically depends on a specific choice of representation. To factor out this dependency in the representation and in respective algorithms gives rise to computationally challenging problems. Determining point correspondences, for example, becomes particularly difficult for shapes in higher dimensions (e.g., surfaces in 3D).
- In particular, the evolution of explicit shape representations requires sophisticated numerical regridding procedures to assure an equidistant spacing of control points and prevent control point overlap.
- Parametric representations are difficult to adapt to varying topology of the represented shape. Numerically, topology changes require sophisticated splitting and remerging procedures.
- A number of recent publications [11, 49, 59] indicate that in contrast to explicit shape representations, the implicit representation of shape allows to compute *globally optimal* solutions to shape inference for large classes of commonly used energy functionals.

A mathematical representation of shape which is independent of parameterization was pioneered in the analysis of random shapes by Fréchet [45] and in the school of mathematical morphology founded by Matheron and Serra [64, 94]. The level set method [39, 72] provides a means of propagating contours $\mathcal{C}$ (independent of parameterization) by evolving associated embedding functions ϕ via partial differential equations – see Fig. 10 for a visualization of the level set function associated with a human silhouette. It has been adapted to segment images based on numerous low-level criteria such as edge consistency [10, 56, 63], intensity

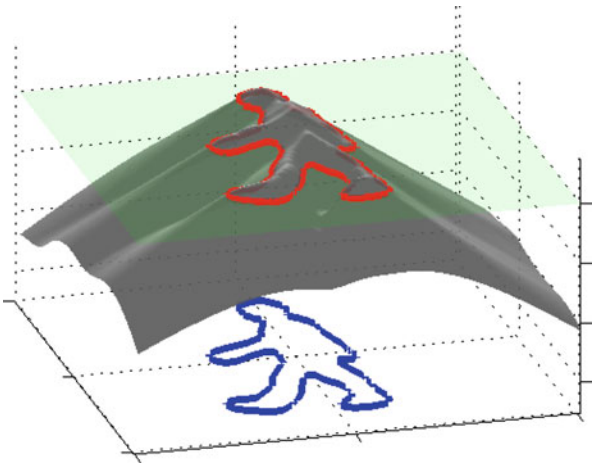


Fig. 10 The level set method is based on representing shapes implicitly as the zero level set of a higher-dimensional embedding function

homogeneity [12, 101], texture information [9, 51, 73, 81], and motion information [33].

In this section, we will give a brief insight into shape modeling and shape priors for implicit level set representations. Parts of the following text were adopted from [34, 35, 82].

Shape Distances for Level Sets

The first step in deriving a shape prior is to define a distance or dissimilarity measure for two shapes encoded by the level set functions ϕ_1 and ϕ_2. We shall briefly discuss three solutions to this problem. In order to guarantee a unique correspondence between a given shape and its embedding function ϕ, we will in the following assume that ϕ is a *signed distance function*, i.e., $\phi > 0$ inside the shape, $\phi < 0$ outside, and $|\nabla\phi| = 1$ almost everywhere. A method to project a given embedding function onto the space of signed distance functions was introduced in [98].

Given two shapes encoded by their signed distance functions ϕ_1 and ϕ_2, a simple measure of their dissimilarity is given by their L_2-distance in Ω [62]:

$$\int_{\Omega} (\phi_1 - \phi_2)^2 \, dx. \tag{17}$$

This measure has the drawback that it depends on the domain of integration Ω. The shape dissimilarity will generally grow if the image domain is increased – even if the relative position of the two shapes remains the same. Various remedies to this problem have been proposed. We refer to [32] for a detailed discussion.

An alternative dissimilarity measure between two implicitly represented shapes represented by the embedding functions ϕ_1 and ϕ_2 is given by the area of the

symmetric difference [14, 15, 77]:

$$d^2(\phi_1, \phi_2) = \int_\Omega (H\phi_1(x) - H\phi_2(x))^2 \, dx. \tag{18}$$

In the present work, we will define the distance between two shapes based on this measure, because it has several favorable properties. Beyond being independent of the image size Ω, measure (18) defines a distance on the set of shapes: it is nonnegative, symmetric, and fulfills the triangle inequality. Moreover, it is more consistent with the philosophy of the level set method in that it only depends on the *sign* of the embedding function. In practice, this means that one does not need to constrain the two level set functions to the space of signed distance functions. It can be shown [15] that L^∞ and $W^{1,2}$ norms on the signed distance functions induce equivalent topologies as the metric (18).

Since the distance (18) is not differentiable, we will in practice consider an approximation of the Heaviside function H by a smooth (differentiable) version H_ϵ. Moreover, we will only consider gradients of energies with respect to the L_2 norm on the level set function, because they are easy to compute and because variations in the signed distance function correspond to respective variations of the implicitly represented curve. In general, however, these do not coincide with the so-called shape gradients – see [46] for a recent work on this topic.

Invariance by Intrinsic Alignment

One can make use of the shape distance (18) in a segmentation process by adding it as a shape prior $E_{\text{shape}}(\phi) = d^2(\phi, \phi_0)$ in a weighted sum to the data term, which we will assume to be the two-phase version of (3) introduced in [13]:

$$E_{\text{data}}(\phi) = \int_\Omega (I - u_+)^2 H\phi(x) dx + \int_\Omega (I - u_-)^2 (1 - H\phi(x)) \, dx + \nu \int_\Omega |\nabla H\phi| dx, \tag{19}$$

Minimizing the total energy

$$E_{\text{total}}(\phi) = E_{\text{data}}(\phi) + \alpha \, E_{\text{shape}}(\phi) = E_{\text{data}}(\phi) + \alpha \, d^2(\phi, \phi_0), \tag{20}$$

with a weight $\alpha > 0$, induces an additional driving term which aims at maximizing the similarity of the evolving shape with a given template shape encoded by the function ϕ_0.

By construction this shape prior is not invariant with respect to certain transformations such as translation, rotation, and scaling of the shape represented by ϕ.

A common approach to introduce invariance (cf. [17, 35, 83]) is to enhance the prior by a set of explicit parameters to account for translation by μ, rotation by an angle θ, and scaling by σ of the shape:

$$d^2(\phi, \phi_0, \mu, \theta, \sigma) = \int_\Omega \left(H\left(\phi\left(\sigma R_\theta(x-\mu)\right)\right) - H\phi_0(x)\right)^2 \, dx. \tag{21}$$

This approach to estimate the appropriate transformation parameters has several drawbacks:

- Optimization of the shape energy (21) is done by local gradient descent. In particular, this implies that one needs to determine an appropriate time step for each parameter, chosen so as to guarantee stability of resulting evolution. In numerical experiments, we found that balancing these parameters requires a careful tuning process.
- The optimization of μ, θ, σ, and ϕ is done simultaneously. In practice, however, it is unclear how to alternate between the updates of the respective parameters. How often should one iterate one or the other gradient descent equation? In experiments, we found that the final solution depends on the selected scheme of optimization.
- The optimal values for the transformation parameters will depend on the embedding function ϕ. An accurate shape gradient should therefore take into account this dependency. In other words, the gradient of (21) with respect to ϕ should take into account how the optimal transformation parameters $\mu(\phi)$, $\sigma(\phi)$, and $\theta(\phi)$ vary with ϕ.

Inspired by the normalization for explicit representations introducing in (10), we can eliminate these difficulties associated with the local optimization of explicit transformation parameters by introducing an intrinsic registration process. We will detail this for the cases of translation and scaling. Extensions to rotation and other transformations are conceivable but will not be pursued here.

Translation Invariance by Intrinsic Alignment

Assume that the template shape represented by ϕ_0 is aligned with respect to the shape's centroid. Then we define a shape energy by

$$E_{\text{shape}}(\phi) = d^2(\phi, \phi_0) = \int_\Omega \left(H\phi(x+\mu_\phi) - H\phi_0(x)\right)^2 \, dx, \tag{22}$$

where the function ϕ is evaluated in coordinates relative to its center of gravity μ_ϕ given by

$$\mu_\phi = \int x\, h\phi \, dx, \quad \text{with } h\phi \equiv \frac{H\phi}{\int_\Omega H\phi \, dx}. \tag{23}$$

This intrinsic alignment guarantees that the distance (22) is invariant to the location of the shape ϕ. In contrast to the shape energy (21), we no longer need to iteratively update an estimate of the location of the object of interest.

Translation and Scale Invariance via Alignment

Given a template shape (represented by ϕ_0) which is normalized with respect to translation and scaling, one can extend the above approach to a shape energy which is invariant to translation and scaling:

$$E_{\text{shape}}(\phi) = d^2(\phi, \phi_0) = \int_\Omega \left(H\phi(\sigma_\phi\, x + \mu_\phi) - H\phi_0(x)\right)^2 dx, \tag{24}$$

where the level set function ϕ is evaluated in coordinates relative to its center of gravity μ_ϕ and in units given by its intrinsic scale σ_ϕ defined as

$$\sigma_\phi = \left(\int (x-\mu)^2\, h\phi\, dx\right)^{\frac{1}{2}}, \quad \text{where } h\phi = \frac{H\phi}{\int_\Omega H\phi\, dx}. \tag{25}$$

In the following, we will show that functional (24) is invariant with respect to translation and scaling of the shape represented by ϕ. Let ϕ be a level set function representing a shape which is centered and normalized such that $\mu_\phi = 0$ and $\sigma_\phi = 1$. Let $\tilde{\phi}$ be an (arbitrary) level set function encoding the same shape after scaling by $\sigma \in \mathbb{R}$ and shifting by $\mu \in \mathbb{R}^2$:

$$H\tilde{\phi}(x) = H\phi\left(\frac{x-\mu}{\sigma}\right).$$

Indeed, center and intrinsic scale of the transformed shape are given by

$$\begin{aligned}
\mu_{\tilde{\phi}} &= \frac{\int x H\tilde{\phi} dx}{\int H\tilde{\phi} dx} = \frac{\int x H\phi\left(\frac{x-\mu}{\sigma}\right) dx}{\int H\phi\left(\frac{x-\mu}{\sigma}\right) dx} = \frac{\int (\sigma x' + \mu) H\phi(x')\, \sigma dx'}{\int H\phi(x')\, \sigma dx'} \\
&= \sigma\mu_\phi + \mu = \mu, \\
\sigma_{\tilde{\phi}} &= \left(\frac{\int (x-\mu_{\tilde{\phi}})^2 H\tilde{\phi} dx}{\int H\tilde{\phi} dx}\right)^{\frac{1}{2}} = \left(\frac{\int (x-\mu)^2 H\phi\left(\frac{x-\mu}{\sigma}\right) dx}{\int H\phi\left(\frac{x-\mu}{\sigma}\right) dx}\right)^{\frac{1}{2}} \\
&= \left(\frac{\int (\sigma x')^2 H\phi(x') dx'}{\int H\phi(x') dx'}\right)^{\frac{1}{2}} = \sigma.
\end{aligned}$$

The shape energy (21) evaluated for $\tilde{\phi}$ is given by

$$\begin{aligned}
E_{\text{shape}}(\tilde{\phi}) &= \int_\Omega \left(H\tilde{\phi}(\sigma_{\tilde{\phi}}\, x + \mu_{\tilde{\phi}}) - H\phi_0(x)\right)^2 dx \\
&= \int_\Omega \left(H\tilde{\phi}(\sigma\, x + \mu) - H\phi_0(x)\right)^2 dx \\
&= \int_\Omega \left(H\phi(x) - H\phi_0(x)\right)^2 dx \;=\; E_{\text{shape}}(\phi).
\end{aligned}$$

Therefore, the above shape dissimilarity measure is invariant with respect to translation and scaling.

Note, however, that while this analytical solution guarantees an invariant shape distance, the transformation parameters μ_ϕ and σ_ϕ are not necessarily the ones which minimize the shape distance (21). Extensions of this approach to a larger class of invariance are conceivable. For example, one could generate invariance with respect to rotation by rotational alignment with respect to the (oriented) principal axis of the shape encoded by ϕ. We will not pursue this here.

Kernel Density Estimation in the Level Set Domain

In the previous sections, we have introduced a translation and scale invariant shape energy and demonstrated its effect on the reconstruction of a corrupted version of a single familiar silhouette the pose of which was unknown. In many practical problems, however, we do not have the exact silhouette of the object of interest. There may be several reasons for this:

- The object of interest may be three-dimensional. Rather than trying to reconstruct the three-dimensional object (which generally requires multiple images and the estimation of correspondence), one may learn the two-dimensional appearance from a set of sample views. A meaningful shape dissimilarity measure should then measure the dissimilarity with respect to this set of projections – see the example in Fig. 8.
- The object of interest may be one object out of a class of similar objects (the class of cars or the class of tree leafs). Given a limited number of training shapes sampled from the class, a useful shape energy should provide the dissimilarity of a particular silhouette with respect to this class.
- Even a single object, observed from a single viewpoint, may exhibit strong shape deformation – the deformation of a gesticulating hand or the deformation which a human silhouette undergoes while walking. In the following, we will assume that one can merely generate a set of stills corresponding to various (randomly sampled) views of the object of interest for different deformations – see Fig. 11. In the following, we will demonstrate that – without constructing a dynamical model of the walking process – one can exploit this set of sample views in order to improve the segmentation of a walking person.

In the above cases, the construction of appropriate shape dissimilarity measures amounts to a problem of density estimation. In the case of explicitly represented boundaries, this has been addressed by modeling the space of familiar shapes by linear subspaces (PCA) [22] and the related Gaussian distribution [36], by mixture models [21] or nonlinear (multimodal) representations via simple models in appropriate feature spaces [27, 28].

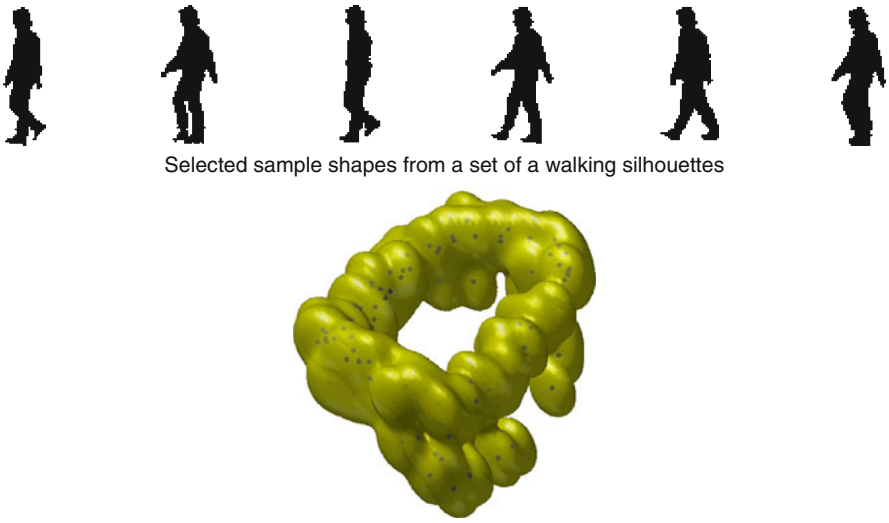


Fig. 11 Density estimated using a kernel density estimator for a projection of 100 silhouettes of a walking person (see above) onto the first three principal components

For level set-based shape representations, it was suggested [62, 84, 100] to fit a linear subspace to the sampled signed distance functions. Alternatively, it was suggested to represent familiar shapes by the level set function encoding the mean shape and a (spatially independent) Gaussian fluctuation at each image location [83]. These approaches were shown to capture some shape variability. Yet, they exhibit two limitations: Firstly, they rely on the assumption of a Gaussian distribution which is not well suited to approximate shape distributions encoding more complex shape variation. Secondly, they work under the assumption that shapes are represented by signed distance functions. Yet, the space of signed distance functions is not a linear space. Therefore, in general, neither the mean nor the linear combination of a set of signed distance functions will correspond to a signed distance function.

In the following, we will propose an alternative approach to generate a statistical shape dissimilarity measure for level set based shape representations. It is based on classical methods of (so-called nonparametric) kernel density estimation and overcomes the above limitations.

Given a set of training shapes $\{\phi_i\}_{i=1\ldots N}$ – such as those shown in Fig. 11 – we define a probability density on the space of signed distance functions by integrating the shape distances (22) or (24) in a Parzen–Rosenblatt kernel density estimator [75, 79]:

$$\mathcal{P}(\phi) \propto \frac{1}{N} \sum_{i=1}^{N} \exp\left(-\frac{1}{2\sigma^2}\, d^2(H\phi, H\phi_i)\right). \tag{26}$$

The kernel density estimator is among the theoretically most studied density estimation methods. It was shown (under fairly mild assumptions) to converge to the true distribution in the limit of infinite samples (and $\sigma \to 0$); the asymptotic convergence rate was studied for different choices of kernel functions.

It should be pointed out that the theory of classical nonparametric density estimation was developed for the case of finite-dimensional data. It is beyond the scope of this work to develop a general theory of probability distributions and density estimation on infinite-dimensional spaces (including issues of integrability and measurable sets). For a general formalism to model probability densities on infinite-dimensional spaces, we refer the reader to the theory of Gaussian processes [76]. In our case, an extension to infinite-dimensional objects such as level set surfaces $\phi : \Omega \to \mathbb{R}$ could be tackled by considering discrete (finite-dimensional) approximations $\{\phi_{ij} \in \mathbb{R}\}_{i=1,\ldots,N, j=1,\ldots,M}$ of these surfaces at increasing levels of spatial resolution and studying the limit of infinitesimal grid size (i.e., $N, M \to \infty$). Alternatively, given a finite number of samples, one can apply classical density estimation techniques efficiently in the finite-dimensional subspace spanned by the training data [82].

Similarly respective metrics on the space of curves give rise to different kinds of gradient descent flows. Recently researchers have developed rather sophisticated metrics to favor smooth transformations or rigid body motions. We refer the reader to [16,97] for promising advances in this direction. In the following we will typically limit ourselves to L_2 gradients.

There exist extensive studies on how to optimally choose the kernel width σ based on asymptotic expansions such as the parametric method [37], heuristic estimates [95, 104], or maximum likelihood optimization by cross validation [18, 42]. We refer to [40, 96] for a detailed discussion. For this work, we simply fix σ^2 to be the mean squared nearest-neighbor distance:

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} \min_{j \neq i} d^2(H\phi_i, H\phi_j). \tag{27}$$

The intuition behind this choice is that the width of the Gaussians is chosen such that on the average the next training shape is within one standard deviation.

Reverting to kernel density estimation resolves the drawbacks of existing approaches to shape models for level set segmentation discussed above. In particular:

- The silhouettes of a rigid 3D object or a deformable object with few degrees of freedom can be expected to form fairly low-dimensional manifolds. The kernel density estimator can capture these without imposing the restrictive assumption of a Gaussian distribution. Figure 11 shows a 3D approximation of our method: We simply projected the embedding functions of 100 silhouettes of a walking person onto the first three eigenmodes of the distribution. The projected silhouette data and the kernel density estimate computed in the 3D

subspace indicate that the underlying distribution is not Gaussian. The estimated distribution (indicated by an isosurface) shows a closed loop which stems from the fact that the silhouettes were drawn from an essentially periodic process.

- Kernel density estimators were shown to converge to the true distribution in the limit of infinite (independent and identically distributed) training samples [40, 96]. In the context of shape representations, this implies that our approach is capable of accurately representing arbitrarily complex shape deformations.
- By not imposing a linear subspace, we circumvent the problem that the space of shapes (and signed distance functions) is not a linear space. In other words, Kernel density estimation allows to estimate distributions on nonlinear (curved) manifolds. In the limit of infinite samples and kernel width σ going to zero, the estimated distribution is more and more constrained to the manifold defined by the shapes.

Gradient Descent Evolution for the Kernel Density Estimator

In the following, we will detail how the statistical distribution (26) can be used to enhance level set based segmentation process. As for the case of parametric curves, we formulate level set segmentation as a problem of Bayesian inference, where the segmentation is obtained by maximizing the conditional probability:

$$\mathcal{P}(\phi \mid I) = \frac{\mathcal{P}(I \mid \phi)\, \mathcal{P}(\phi)}{\mathcal{P}(I)}, \tag{28}$$

with respect to the level set function ϕ, given the input image I. For a given image, this is equivalent to minimizing the negative log-likelihood which is given by a sum of two energies:

$$E(\phi) = E_{\text{data}}(\phi) \; + \; E_{\text{shape}}(\phi), \tag{29}$$

with

$$E_{\text{shape}}(\phi) = -\log \mathcal{P}(\phi). \tag{30}$$

Minimizing the energy (29) generates a segmentation process which simultaneously aims at maximizing intensity homogeneity in the separated phases and a similarity of the evolving shape with respect to all the training shapes encoded through the statistical estimator (26).

Gradient descent with respect to the embedding function amounts to the evolution:

$$\frac{\partial \phi}{\partial t} = -\frac{1}{\alpha} \frac{\partial E_{\text{data}}}{\partial \phi} - \frac{\partial E_{\text{shape}}}{\partial \phi}, \tag{31}$$

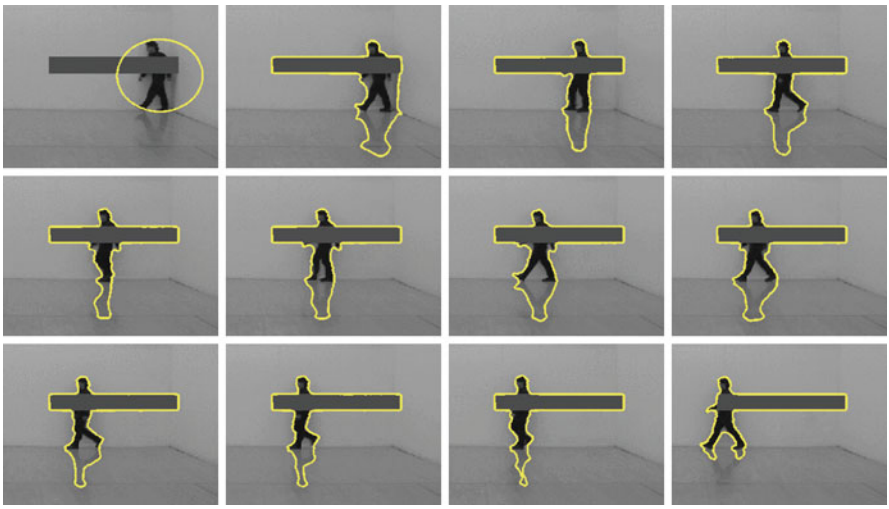


Fig. 12 Purely intensity-based segmentation. Various frames show the segmentation of a partially occluded walking person generated by minimizing the Chan–Vese energy (19). The walking person cannot be separated from the occlusion and darker areas of the background such as the person's shadow

where the knowledge-driven component is given by

$$\frac{\partial E_{\text{shape}}}{\partial \phi} = \frac{\sum \alpha_i \frac{\partial}{\partial \phi} d^2(H\phi, H\phi_i)}{2\sigma^2 \sum \alpha_i}, \tag{32}$$

which simply induces a force in direction of each training shape ϕ_i weighted by the factor:

$$\alpha_i = \exp\left(-\frac{1}{2\sigma^2} d^2(H\phi, H\phi_i)\right), \tag{33}$$

which decays exponentially with the distance from the training shape ϕ_i.

Nonlinear Shape Priors for Tracking a Walking Person

In the following, we apply the above shape prior to the segmentation of a partially occluded walking person. To this end, a sequence of a walking figure was partially occluded by an artificial bar. Subsequently we minimized energy (19), segmenting each frame of the sequence using the previous segmentation as initialization. Figure 12 shows that this purely image-driven segmentation scheme is not capable of separating the object of interest from the occluding bar and similarly shaded background regions such as the object's shadow on the floor.

In a second experiment, we manually binarized the images corresponding to the first half of the original sequence (frames 1 through 42) and aligned them to their

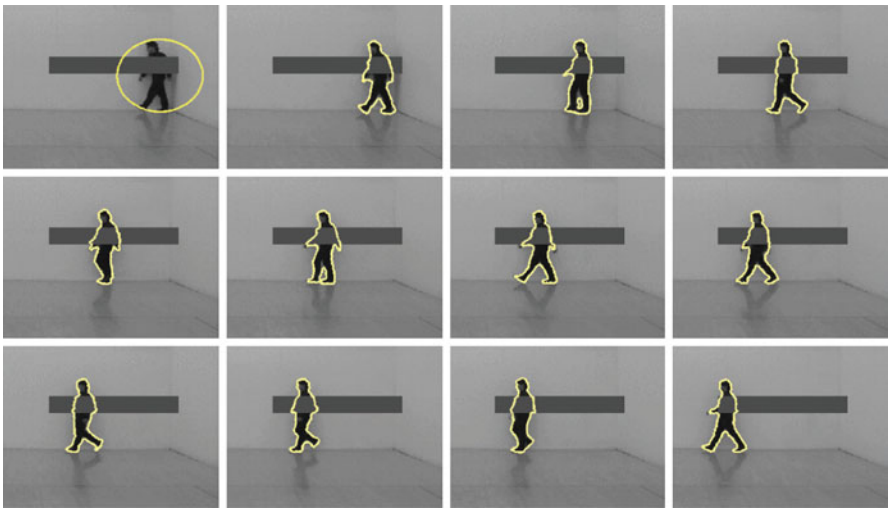


Fig. 13 Segmentation with nonparametric invariant shape prior. Segmentation generated by minimizing energy (29) combining intensity information with the shape prior (26). For every frame in the sequence, the gradient descent equation was iterated (with fixed parameters), using the previous segmentation as initialization. The shape prior permits to separate the walking person from the occlusion and darker areas of the background such as the shadow. The shapes in the second half of the sequence were not part of the training set

respective center of gravity to obtain a set of training shape – see Fig. 11. Then we ran the segmentation process (31) with the shape prior (26). Apart from adding the shape prior, we kept the other parameters constant for comparability.

Figure 13 shows several frames from this knowledge-driven segmentation. A comparison to the corresponding frames in Fig. 12 demonstrates several properties:

- The shape prior permits to accurately reconstruct an entire set of fairly different shapes. Since the shape prior is defined on the level set function ϕ – rather than on the boundary C (cf. [17]) – it can easily handle changing topology.
- The shape prior is invariant to translation such that the object silhouette can be reconstructed in arbitrary locations of the image.
- The statistical nature of the prior allows to also reconstruct silhouettes which were not part of the training set – corresponding to the second half of the images shown (beyond frame 42).

5 Dynamical Shape Priors for Implicit Shapes

Capturing the Temporal Evolution of Shape

In the above works, statistically learned shape information was shown to cope for missing or misleading information in the input images due to noise, clutter, and

occlusion. The shape priors were developed to segment objects of familiar shape in a given image. Although they can be applied to tracking objects in image sequences, they are not well-suited for this task, because they neglect the *temporal coherence of silhouettes* which characterizes many deforming shapes.

When tracking a deformable object over time, clearly not all shapes are equally likely at a given time instance. Regularly sampled images of a walking person, for example, exhibit a typical pattern of consecutive silhouettes. Similarly, the projections of a rigid 3D object rotating at constant speed are generally not independent samples from a statistical shape distribution. Instead, the resulting set of silhouettes can be expected to contain strong temporal correlations.

In the following, we will present temporal statistical shape models for implicitly represented shapes that were first introduced in [26]. In particular, the shape probability at a given time depends on the shapes observed at previous time steps. The integration of such dynamical shape models into the segmentation process can be elegantly formulated within a Bayesian framework for level set based image sequence segmentation. The resulting optimization by gradient descent induces an evolution of the level set function which is driven both by the intensity information of the current image as well as by a dynamical shape prior which relies on the segmentations obtained on the preceding frames. Experimental evaluation demonstrates that the resulting segmentations are not only similar to previously learned shapes, but they are also consistent with the temporal correlations estimated from sample sequences. The resulting segmentation process can cope with large amounts of noise and occlusion because it exploits prior knowledge about *temporal* shape consistency and because it aggregates information from the input images over time (rather than treating each image independently).

Level Set-Based Tracking via Bayesian Inference

Statistical models can be estimated more reliably if the dimensionality of the model and the data are low. We will therefore cast the Bayesian inference in a low-dimensional formulation within the subspace spanned by the largest principal eigenmodes of a set of sample shapes. We exploit the training sequence in a twofold way: Firstly, it serves to define a low-dimensional subspace in which to perform estimation. And secondly, within this subspace we use it to learn dynamical models for implicit shapes. For static shape priors this concept was already used in [82].

Let $\{\phi_1, \ldots, \phi_N\}$ be a temporal sequence of training shapes. (We assume that all training shapes ϕ_i are signed distance functions. Yet an arbitrary linear combination of eigenmodes will in general not generate a signed distance function. While the discussed statistical shape models favor shapes which are close to the training shapes (and therefore close to the set of signed distance functions), not all shapes sampled in the considered subspace will correspond to signed distance functions.) Let ϕ_0 denote the mean shape and $\psi_1, \ldots, \psi_n$ the n largest eigenmodes with $n \ll N$. We will then approximate each training shape as

$$\phi_i(x) = \phi_0(x) + \sum_{j=1}^{n} \alpha_{ij}\, \psi_j(x), \tag{34}$$

where

$$\alpha_{ij} = \langle \phi_i - \phi_0, \psi_j \rangle \equiv \int (\phi_i - \phi_0)\, \psi_j \, dx. \tag{35}$$

Such PCA-based representations of level set functions have been successfully applied for the construction of statistical shape priors in [62, 82, 84, 100]. In the following, we will denote the vector of the first n eigenmodes as $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_n)$. Each sample shape ϕ_i is therefore approximated by the n-dimensional shape vector $\boldsymbol{\alpha}_i = (\alpha_{i1}, \ldots, \alpha_{in})$. Similarly, an arbitrary shape ϕ can be approximated by a shape vector of the form

$$\boldsymbol{\alpha}_\phi = \langle \phi - \phi_0, \boldsymbol{\psi} \rangle. \tag{36}$$

In addition to the deformation parameters $\boldsymbol{\alpha}$, we introduce transformation parameters θ, and we introduce the notation:

$$\phi_{\boldsymbol{\alpha},\theta}(x) = \phi_0(T_\theta x) + \boldsymbol{\alpha}^\top \boldsymbol{\psi}(T_\theta x), \tag{37}$$

to denote the embedding function of a shape generated with deformation parameters $\boldsymbol{\alpha}$ and transformed with parameters θ. The transformations T_θ can be translation, rotation, and scaling (depending on the application).

With this notation, the goal of image sequence segmentation within this subspace can be stated as follows: Given consecutive images $I_t : \Omega \to \mathbb{R}$ from an image sequence and given the segmentations $\hat{\boldsymbol{\alpha}}_{1:t-1}$ and transformations $\hat{\theta}_{1:t-1}$ obtained on the previous images $I_{1:t-1}$, compute the most likely deformation $\hat{\boldsymbol{\alpha}}_t$ and transformation $\hat{\theta}_t$ by maximizing the conditional probability:

$$\mathcal{P}(\boldsymbol{\alpha}_t, \theta_t \mid I_t, \hat{\boldsymbol{\alpha}}_{1:t-1}, \hat{\theta}_{1:t-1}) = \frac{\mathcal{P}(I_t \mid \boldsymbol{\alpha}_t, \theta_t)\, \mathcal{P}(\boldsymbol{\alpha}_t, \theta_t \mid \hat{\boldsymbol{\alpha}}_{1:t-1}, \hat{\theta}_{1:t-1})}{\mathcal{P}(I_t \mid \hat{\boldsymbol{\alpha}}_{1:t-1}, \hat{\theta}_{1:t-1})}. \tag{38}$$

The key challenge, addressed in the following, is to model the conditional probability:

$$\mathcal{P}(\boldsymbol{\alpha}_t, \theta_t \mid \hat{\boldsymbol{\alpha}}_{1:t-1}, \hat{\theta}_{1:t-1}), \tag{39}$$

which constitutes the probability for observing a particular shape $\boldsymbol{\alpha}_t$ and a particular transformation θ_t at time t, conditioned on the parameter estimates for shape and transformation obtained on previous images.

Linear Dynamical Models for Implicit Shapes

For realistic deformable objects, one can expect the deformation parameters $\boldsymbol{\alpha}_t$ and the transformation parameters θ_t to be tightly coupled. Yet, we want to learn dynamical shape models which are invariant to the absolute translation, rotation, etc. To this end, we can make use of the fact that the transformations form a group which implies that the transformation θ_t at time t can be obtained from the previous transformation θ_{t-1} by applying an incremental transformation $\triangle\theta_t$: $T_{\theta_t} x = T_{\triangle\theta_t} T_{\theta_{t-1}} x$. Instead of learning models of the absolute transformation θ_t, we can simply learn models of the update transformations $\triangle\theta_t$ (e.g., the changes in translation and rotation). By construction, such models are invariant with respect to the global pose or location of the modeled shape.

To jointly model transformation and deformation, we simply obtain for each training shape in the learning sequence the deformation parameters $\boldsymbol{\alpha}_i$ and the transformation changes $\triangle\theta_i$ and define the *extended shape vector*:

$$\boldsymbol{\beta}_t := \begin{pmatrix} \boldsymbol{\alpha}_t \\ \triangle\theta_t \end{pmatrix}. \tag{40}$$

We will then impose a linear dynamical model of order k to approximate the temporal evolution of the extended shape vector:

$$\mathcal{P}(\boldsymbol{\beta}_t \,|\, \hat{\boldsymbol{\beta}}_{1:t-1}) \propto \exp\left(-\frac{1}{2}\, \boldsymbol{\nu}^\top \, \Sigma^{-1} \, \boldsymbol{\nu}\right), \tag{41}$$

where

$$\boldsymbol{\nu} \equiv \boldsymbol{\beta}_t - \boldsymbol{\mu} - A_1 \hat{\boldsymbol{\beta}}_{t-1} - A_2 \hat{\boldsymbol{\beta}}_{t-2} \ldots - A_k \hat{\boldsymbol{\beta}}_{t-k}. \tag{42}$$

Various methods have been proposed in the literature to estimate the model parameters given by the mean $\boldsymbol{\mu}$ and the transition and noise matrices $A_1, \ldots, A_k, \Sigma$. We applied a stepwise least squares algorithm proposed in [71]. Using dynamical models up to an order of 8, we found that according to Schwarz's Bayesian criterion [92], our training sequences were best approximated by an autoregressive model of second order ($k = 2$).

Figure 14 shows a sequence of statistically synthesized embedding functions and the induced contours given by the zero level line of the respective surfaces – for easier visualization, the transformational degrees are neglected. In particular, the implicit representation allows to synthesize shapes of varying topology. The silhouette on the bottom left of Fig. 14, for example, consists of two contours.

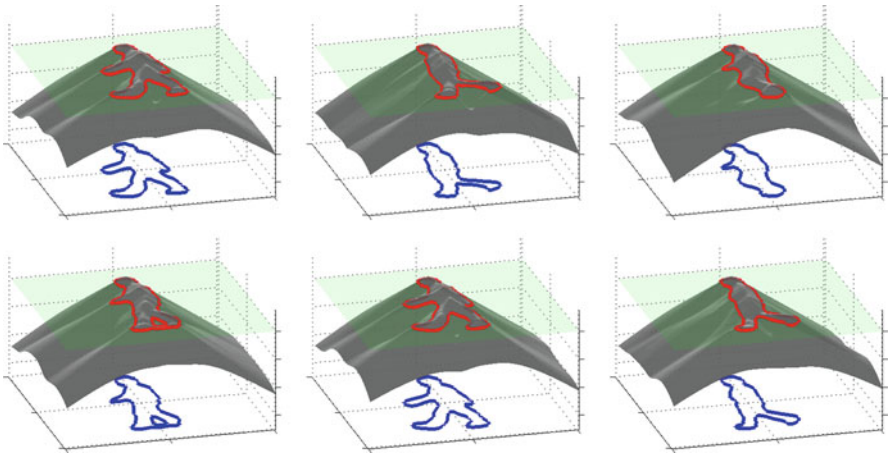


Fig. 14 Synthesis of implicit dynamical shapes. Statistically generated embedding surfaces obtained by sampling from a second-order autoregressive model and the contours given by the zero level lines of the synthesized surfaces. The implicit representation allows the embedded contour to change topology (*bottom left image*)

Variational Segmentation with Dynamical Shape Priors

Given an image I_t from an image sequence and given a set of previously segmented shapes with shape parameters $\hat{\boldsymbol{\alpha}}_{1:t-1}$ and transformation parameters $\hat{\theta}_{1:t-1}$, the goal of tracking is to maximize the conditional probability (38) with respect to shape $\boldsymbol{\alpha}_t$ and transformation θ_t. This can be performed by minimizing its negative logarithm, which is – up to a constant – given by an energy of the form

$$E(\boldsymbol{\alpha}_t, \theta_t) = E_{\text{data}}(\boldsymbol{\alpha}_t, \theta_t) + E_{\text{shape}}(\boldsymbol{\alpha}_t, \theta_t). \tag{43}$$

For the data term we use the model in (3) with independent intensity variances:

$$E_{\text{data}}(\boldsymbol{\alpha}_t, \theta_t) = \int \left(\frac{(I_t - \mu_1)^2}{2\sigma_1^2} + \log \sigma_1 \right) H\phi_{\boldsymbol{\alpha}_t, \theta_t} + \left(\frac{(I_t - \mu_2)^2}{2\sigma_2^2} + \log \sigma_2 \right) (1 - H\phi_{\boldsymbol{\alpha}_t, \theta_t})\, dx. \tag{44}$$

Using the autoregressive model (41), the shape energy is given by

$$E_{\text{shape}}(\boldsymbol{\alpha}_t, \theta_t) = \frac{1}{2}\, \boldsymbol{\nu}^\top\, \Sigma^{-1}\, \boldsymbol{\nu}, \tag{45}$$

with $\boldsymbol{\nu}$ defined in (42).

The total energy (43) is easily minimized by gradient descent. For details we refer to [26].

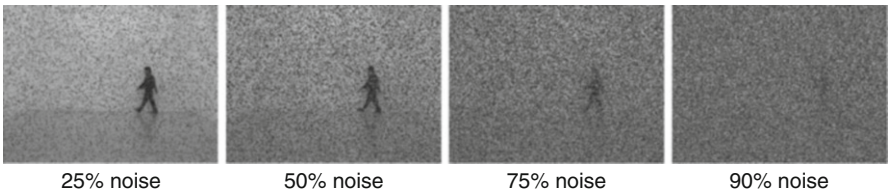


Fig. 15 Images from a sequence with increasing amount of noise

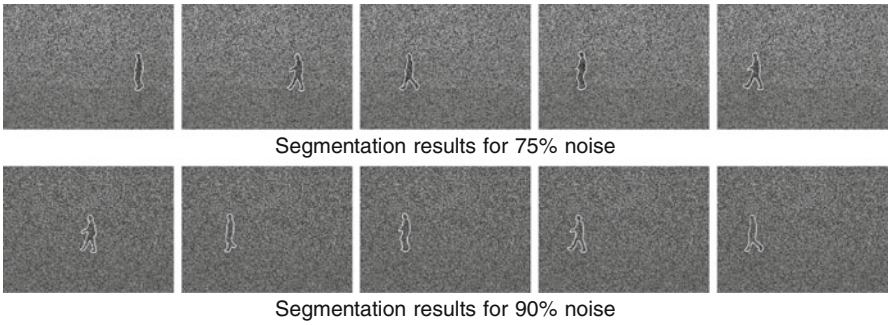


Fig. 16 Variational image sequence segmentation with a dynamical shape prior for various amounts of noise. Ninety percent noise means that nine out of ten intensity values were replaced by a random intensity from a uniform distribution. The statistically learned dynamical model allows for reliable segmentation results despite prominent amounts of noise

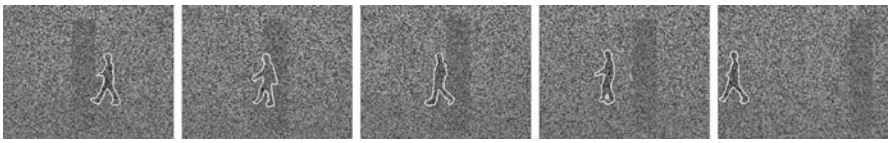


Fig. 17 Tracking in the presence of occlusion. The dynamical shape prior allows to reliably segment the walking person despite noise and occlusion

Figure 15 shows images from a sequence that was degraded by increasing amounts of noise.

Figure 16 shows segmentation results obtained by minimizing (43) as presented above. Despite prominent amounts of noise, the segmentation process provides reliable segmentations where human observers fail.

Figure 17 shows the segmentation of an image sequence showing a walking person that was corrupted by noise and an occlusion which completely covers the walking person for several frames. The dynamical shape prior allows for reliable segmentations despite noise and occlusion. For more details and quantitative evaluations, we refer to [26].

6 Parametric Representations Revisited: Combinatorial Solutions for Segmentation with Shape Priors

In previous sections we saw that shape priors allow to improve the segmentation and tracking of familiar deformable objects, biasing the segmentation process to favor familiar shapes or even familiar shape evolution. Unfortunately, these approaches are based on locally minimizing respective energies via gradient descent. Since these energies are generally non-convex, respective solutions are bound to be locally optimal only. As a consequence, they depend on an initialization and are likely to be suboptimal in practice. One exception based on implicit shape representations as binary indicator functions and convex relaxation techniques was proposed in [31]. Yet, the linear interpolation of shapes represented by binary indicator functions does not give rise to plausible intermediate shapes such that respective algorithms require a large number of training shapes.

Moreover, while implicit representations like the level set method circumvent the problem of computing correspondences between points on either of two shapes, it is well-known that the aspect of point correspondences plays a vital role in human notions of shape similarity. For matching planar shapes there is abundant literature on how to solve the arising correspondence problem in polynomial time using dynamic programming techniques [48, 85, 93].

Similar concepts of dynamic programming can be employed to localize deformed template curves in images. Coughlan et al. [23] detected open boundaries by shortest-path algorithms in higher-dimensional graphs. And Felzenszwalb et al. used dynamic programming in chordal graphs to localize shapes, albeit not on a pixel level.

Polynomial-time solutions for localizing deformable closed template curves in images using minimum ratio cycles or shortest circular paths were proposed in [89], with a further generalization presented in [88]. There the problem of determining a segmentation of an image $I : \Omega \to \mathbb{R}$ that is elastically similar to an observed template $cc : \mathbb{S}^1 \to \mathbb{R}^2$ by computing minimum ratio cycles

$$\Gamma : \mathbb{S}^1 \to \Omega \times \mathbb{S}^1 \tag{46}$$

in the product space spanned by the image domain Ω and template domain $\mathbb{S}^1$. See Fig. 18 for a schematic visualization. All points along this circular path provide a pair of corresponding template point and image pixel. In this manner, the matching of template points to image pixels is equivalent to the estimation of orientation-preserving cyclic paths, which can be solved in polynomial time using dynamic programming techniques such as ratio cycles [86] or shortest circular paths [91].

Figure 19 shows an example result obtained with this approach: The algorithm determines a deformed version (right) of a template curve (left) in an image (center) in globally optimal manner. An initialization is no longer required and the best conceivable solution is determined in polynomial time.

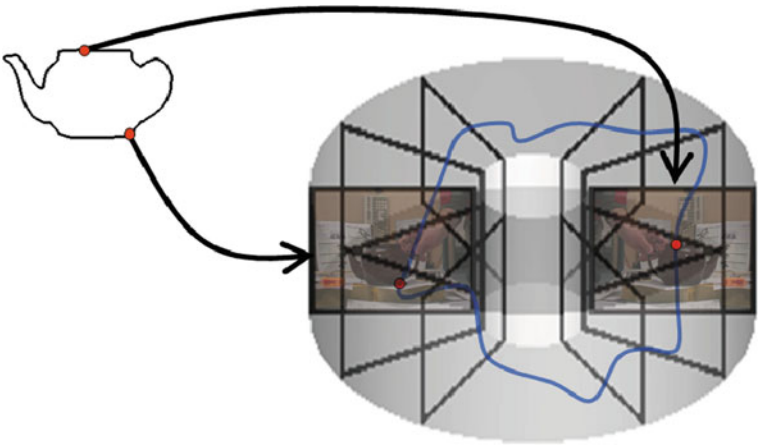


Fig. 18 A polynomial-time solution for matching shapes to images: matching a template curve $C : \mathbb{S}^1 \to \mathbb{R}^2$ (*left*) to the image plane $\Omega \subset \mathbb{R}^2$ is equivalent to computing an orientation-preserving cyclic path $\Gamma : \mathbb{S}^1 \to \Omega \times \mathbb{S}^1$ (*blue curve*) in the product space spanned by the image domain and the template domain. The latter problem can be solved in polynomial time – see [89] for details

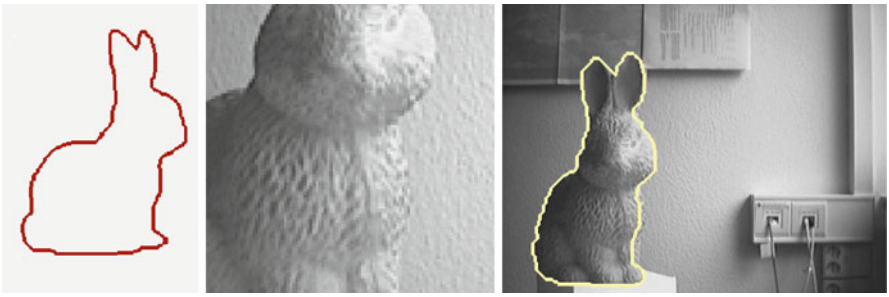


Fig. 19 Segmentation with a single template: despite significant deformation and translation, the initial template curve (*red*) is accurately matched to the low-contrast input image. The globally optimal correspondence between template points and image pixels is computed in polynomial time by dynamic programming techniques [89]

Figure 20 shows further examples of tracking objects: over long sequences of hundreds of frames, the objects of interest are tracked reliably – despite low contrast, camera shake, bad visibility, and illumination changes. For further details we refer to [89].

7 Conclusion

In the previous sections, we have discussed various ways to impose statistical shape priors into image segmentation methods. We have made several observations:

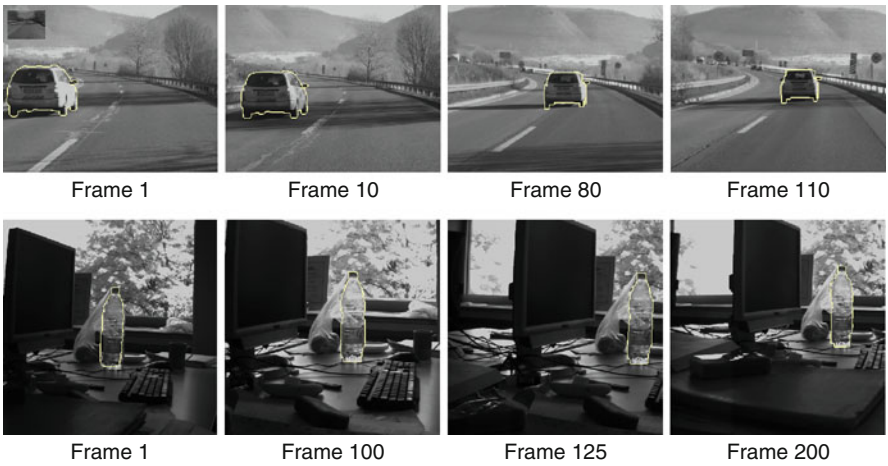


Fig. 20 Tracking of various objects in challenging real-world sequences [89]. Despite bad visibility, camera shake, and substantial lighting changes, the polynomial-time algorithm allows to reliably track objects over hundreds of frames (Image data taken from [89])

- By imposing statistically learned shape information, one can generate segmentation processes which favor the emergence of familiar shapes – where familiarity is based on one or several training shapes.
- Statistical shape information can be elegantly combined with the input image data in the framework of Bayesian maximum a posteriori estimation. Maximizing the posterior distribution is equivalent to minimizing a sum of two energies representing the data term and the shape prior. A further generalization allows to impose dynamical shape priors so as to favor familiar deformations of shape in image sequence segmentation.
- While linear Gaussian shape priors are quite popular, the silhouettes of typical objects in our environment are generally not Gaussian distributed. In contrast to linear Gaussian priors, nonlinear statistical shape priors based on Parzen–Rosenblatt kernel density estimators or based on Gaussian distributions in appropriate feature spaces [28] allow to encode a large variety of rather distinct shapes in a single shape energy.
- Shapes can be represented explicitly (as points on the object's boundary or surface) or implicitly (as the indicator function of the interior of the object). They can be represented in a spatially discrete or a spatially continuous setting.
- The choice of shape representation has important consequences regarding the question which optimization algorithms are employed and whether respective energies can be minimized locally or globally. Moreover, different shape representations give rise to different notions of shape similarity and shape interpolation. As a result, there is no single ideal representation of shape. Ultimately one may favor hybrid representations such as the one proposed in [90]. It combines explicit and implicit representations allowing cost functions which

represent properties of both the object's interior and its boundary. Subsequent LP relaxation provides minimizers of bounded optimality.

Cross-References

- ▸ Level Set Methods for Structural Inversion and Image Reconstruction
- ▸ Manifold Intrinsic Similarity
- ▸ Shape Spaces
- ▸ Variational Methods in Shape Analysis

References

1. Amini, A.A., Weymouth, T.E., Jain, R.C.: Using dynamic programming for solving variational problems in vision. IEEE Trans. Pattern Anal. Mach. Intell. **12**(9), 855–867 (1990)
2. Awate, S.P., Tasdizen, T., Whitaker, R.T.: Unsupervised texture segmentation with non-parametric neighborhood statistics. In: European Conference on Computer Vision (ECCV), pp. 494–507, Graz. Springer (2006)
3. Blake, A., Isard, M.: Active Contours. Springer, London (1998)
4. Blake, A., Zisserman, A.: Visual Reconstruction. MIT, Cambridge (1987)
5. Bookstein, F.L.: The Measurement of Biological Shape and Shape Change. Lecture Notes in Biomath, vol, 24. Springer, New York (1978)
6. Boykov, Y., Kolmogorov, V.: Computing geodesics and minimal surfaces via graph cuts. In: IEEE International Conference on Computer Vision, Nice, pp. 26–33 (2003)
7. Boykov, Y., Kolmogorov, V.: An experimental comparison of min-cut/max-ow algorithms for energy minimization in vision. IEEE Trans. Pattern Anal. Mach. Intell. **26**(9), 1124–1137 (2004)
8. Brox, T., Rousson, M., Deriche, R., Weickert, J.: Unsupervised segmentation incorporating colour, texture, and motion. In: Petkov, N., Westenberg, M.A. (eds.) Computer Analysis of Images and Patterns. LNCS, vol. 2756, pp. 353–360. Springer, Groningen (2003)
9. Brox, T., Weickert, J.: A TV flow based local scale measure for texture discrimination. In: Pajdla, T., Hlavac, V. (eds.) European Conference on Computer Vision. LNCS, vol. 3022, pp. 578–590. Springer, Prague (2004)
10. Caselles, V., Kimmel, R., Sapiro, G.: Geodesic active contours. In: Proceedings of the IEEE International Conference on Computer Vision, Boston, pp. 694–699 (1995)
11. Chan, T., Esedoglu, S., Nikolova, M.: Algorithms for finding global minimizers of image segmentation and denoising models. SIAM J. Appl. Math. **66**(5), 1632–1648 (2006)
12. Chan, T.F., Vese, L.A.: Active contours without edges. IEEE Trans. Image Process. **10**(2), 266–277 (2001)
13. Chan, T.F., Vese, L.A.: A level set algorithm for minimizing the Mumford–Shah functional in image processing. In: IEEE Workshop on Variational and Level Set Methods, Vancouver, pp. 161–168 (2001)
14. Chan, T., Zhu, W.: Level set based shape prior segmentation. Technical report 03-66, Computational Applied Mathematics, UCLA, Los Angeles (2003)
15. Charpiat, G., Faugeras, O., Keriven, R.: Approximations of shape metrics and application to shape warping and empirical shape statistics. J. Found. Comput. Math. **5**(1), 1–58 (2005)
16. Charpiat, G., Faugeras, O., Pons, J.-P., Keriven, R.: Generalized gradients: priors on minimization flows. Int. J. Comput. Vis. **73**(3), 325–344 (2007)

17. Chen, Y., Tagare, H., Thiruvenkadam, S., Huang, F., Wilson, D., Gopinath, K.S., Briggs, R.W., Geiser, E.: Using shape priors in geometric active contours in a variational framework. Int. J. Comput. Vis. **50**(3), 315–328 (2002)
18. Chow, Y.S., Geman, S., Wu, L.D.: Consistent cross-validated density estimation. Ann. Stat. **11**, 25–38 (1983)
19. Cipolla, R., Blake, A.: The dynamic analysis of apparent contours. In: IEEE International Conference on Computer Vision, Osaka, pp. 616–625. Springer (1990)
20. Cohen, L., Kimmel, R.: Global minimum for active contour models: a minimal path approach. Int. J. Comput. Vis. **24**(1), 57–78 (1997)
21. Cootes, T.F., Taylor, C.J.: A mixture model for representing shape variation. Image Vis. Comput. **17**(8), 567–574 (1999)
22. Cootes, T.F., Taylor, C.J., Cooper, D.M., Graham, J.: Active shape models – their training and application. Comput. Vis. Image Underst. **61**(1), 38–59 (1995)
23. Coughlan, J., Yuille, A., English, C., Snow, D.: Efficient deformable template detection and localization without user initialization. Comput. Vis. Image Underst. **78**(3), 303–319 (2000)
24. Courant, R., Hilbert, D.: Methods of Mathematical Physics, vol. 1. Interscience, New York (1953)
25. Cremers, D.: Statistical shape knowledge in variational image segmentation. PhD thesis, Department of Mathematics and Computer Science, University of Mannheim, Germany (2002)
26. Cremers, D.: Dynamical statistical shape priors for level set based tracking. IEEE Trans. Pattern Anal. Mach. Intell. **28**(8), 1262–1273 (2006)
27. Cremers, D., Kohlberger, T., Schnörr, C.: Nonlinear shape statistics in Mumford–Shah based segmentation. In: Heyden, A., et al. (eds.) European Conference on Computer Vision, Copenhagen, pp 93–108. LNCS, vol. 2351. Springer (2002)
28. Cremers, D., Kohlberger, T., Schnörr, C.: Shape statistics in kernel space for variational image segmentation. Pattern Recognit. **36**(9), 1929–1943 (2003)
29. Cremers, D., Osher, S.J., Soatto, S.: Kernel density estimation and intrinsic alignment for shape priors in level set segmentation. Int. J. Comput. Vis. **69**(3), 335–351 (2006)
30. Cremers, D., Rousson, M., Deriche, R.: A review of statistical approaches to level set segmentation: integrating color, texture, motion and shape. Int. J. Comput. Vis. **72**(2), 195–215 (2007)
31. Cremers, D., Schmidt, F.R., Barthel, F.: Shape priors in variational image segmentation: convexity, lipschitz continuity and globally optimal solutions. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Anchorage (2008)
32. Cremers, D., Soatto, S.: A pseudo-distance for shape priors in level set segmentation. In: Paragios, N. (ed.) IEEE 2nd International Workshop on Variational, Geometric and Level Set Methods, Nice, pp. 169–176 (2003)
33. Cremers, D., Soatto, S.: Motion competition: a variational framework for piecewise parametric motion segmentation. Int. J. Comput. Vis. **62**(3), 249–265 (2005)
34. Cremers, D., Sochen, N., Schnörr, C.: A multiphase dynamic labeling model for variational recognition-driven image segmentation. In: Pajdla, T., Hlavac, V. (eds.) European Conference on Computer Vision, Graz. LNCS, vol. 3024, pp 74–86. Springer (2006)
35. Cremers, D., Sochen, N., Schnörr, C.: A multiphase dynamic labeling model for variational recognition-driven image segmentation. Int. J. Comput. Vis. **66**(1), 67–81 (2006)
36. Cremers, D., Tischhäuser, F., Weickert, J., Schnörr, C.: Diffusion snakes: introducing statistical shape knowledge into the Mumford–Shah functional. Int. J. Comput. Vis. **50**(3), 295–313 (2002)
37. Deheuvels, P.: Estimation non paramétrique de la densité par histogrammes généralisés. Revue de Statistique Appliquée **25**, 5–42 (1977)
38. Delingette, H., Montagnat, J.: New algorithms for controlling active contours shape and topology. In: Vernon, D. (ed.) Proceedings of the European Conference on Computer Vision, Dublin. LNCS, vol. 1843, pp. 381–395. Springer (2000)
39. Dervieux, A., Thomasset, F.: A finite element method for the simulation of Raleigh–Taylor instability. Springer Lect. Notes Math. **771**, 145–158 (1979)

40. Devroye, L., Györfi, L.: Nonparametric Density Estimation: The L1 View. Wiley, New York (1985)
41. Dryden, I.L., Mardia, K.V.: Statistical Shape Analysis. Wiley, Chichester (1998)
42. Duin, R.P.W.: On the choice of smoothing parameters for Parzen estimators of probability density functions. IEEE Trans. Comput. **25**, 1175–1179 (1976)
43. Farin, G.: Curves and Surfaces for Computer–Aided Geometric Design. Academic, San Diego (1997)
44. Franchini, E., Morigi, S., Sgallari, F.: Segmentation of 3D tubular structures by a PDE-based anisotropic diffusion model. In: International Conference on Scale Space and Variational Methods, Voss. LNCS, vol. 5567, pp. 75–86. Springer (2009)
45. Fréchet, M.: Les courbes aléatoires. Bull. Int. Stat. Inst. **38**, 499–504 (1961)
46. Fundana, K., Overgaard, N.C., Heyden, A.: Variational segmentation of image sequences using region-based active contours and deformable shape priors. Int. J. Comput. Vis. **80**(3), 289–299 (2008)
47. Gdalyahu, Y., Weinshall, D.: Flexible syntactic matching of curves and its application to automatic hierarchical classication of silhouettes. IEEE Trans. Pattern Anal. Mach. Intell. **21**(12), 1312–1328 (1999)
48. Geiger, D., Gupta, A., Costa, L.A., Vlontzos, J.: Dynamic programming for detecting, tracking and matching deformable contours. IEEE Trans. Pattern Anal. Mach. Intell. **17**(3), 294–302 (1995)
49. Greig, D.M., Porteous, B.T., Seheult, A.H.: Exact maximum a posteriori estimation for binary images. J. R. Stat. Soc. B **51**(2), 271–279 (1989)
50. Grenander, U., Chow, Y., Keenan, D.M.: Hands: A Pattern Theoretic Study of Biological Shapes. Springer, New York (1991)
51. Heiler, M., Schnörr, C.: Natural image statistics for natural image segmentation. In: IEEE International Conference on Computer Vision, Nice, pp. 1259–1266 (2003)
52. Ising, E.: Beitrag zur Theorie des Ferromagnetismus. Zeitschrift für Physik **23**, 253–258 (1925)
53. Kass, M., Witkin, A., Terzopoulos, D.: Snakes: active contour models. Int. J. Comput. Vis. **1**(4), 321–331 (1988)
54. Kendall, D.G.: The diffusion of shape. Adv. Appl. Probab. **9**, 428–430 (1977)
55. Kervrann, C., Heitz, F.: Statistical deformable model-based segmentation of image motion. IEEE Trans. Image Process. **8**, 583–588 (1999)
56. Kichenassamy, S., Kumar, A., Olver, P.J., Tannenbaum, A., Yezzi, A.J.: Gradient ows and geometric active contour models. In: IEEE International Conference on Computer Vision, Cambridge, pp. 810–815 (1995)
57. Kim, J., Fisher, J.W., Yezzi, A., Cetin, M., Willsky, A.: Nonparametric methods for image segmentation using information theory and curve evolution. In: International Conference on Image Processing, Rochester, vol. 3, pp. 797–800 (2002)
58. Kohlberger, T., Cremers, D., Rousson, M., Ramaraj, R.: 4D shape priors for level set segmentation of the left myocardium in SPECT sequences. In: Medical Image Computing and Computer Assisted Intervention. LNCS, vol. 4190, pp. 92–100. Springer, Heidelberg (2006)
59. Kolev, K., Klodt, M., Brox, T., Cremers, D.: Continuous global optimization in multview 3D reconstruction. Int. J. Comput. Vis. **84**, 80–96 (2009)
60. Lachaud, J.-O., Montanvert, A.: Deformable meshes with automated topology changes for coarse-to-fine three-dimensional surface extraction. Med. Image Anal. **3**(2), 187–207 (1999)
61. Leitner, F., Cinquin, P.: Complex topology 3D objects segmentation. In: SPIE Conference on Advances in Intelligent Robotics Systems, Boston, vol. 1609 (1991)
62. Leventon, M., Grimson, W., Faugeras, O.: Statistical shape inuence in geodesic active contours. In: International Conference on Computer Vision and Pattern Recognition, Hilton Head Island, vol. 1, pp. 316–323 (2000)
63. Malladi, R., Sethian, J.A., Vemuri, B.C.: Shape modeling with front propagation: a level set approach. IEEE Trans. Pattern Anal. Mach. Intell. **17**(2), 158–175 (1995)

64. Matheron, G.: Random sets and integral geometry. Wiley, New York (1975)
65. McInerney, T., Terzopoulos, D.: Topologically adaptable snakes. In: Proceedings of the 5th International Conference on Computer Vision, Cambridge, 20–23 June 1995, pp. 840–845. IEEE Computer Society, Los Alamitos (1995)
66. Menet, S., Saint-Marc, P., Medioni, G.: B–snakes: implementation and application to stereo. In: Proceedings of the DARPA Image Understanding Workshop, Pittsburgh, 6–8 Apr 1990, pp. 720–726 (1990)
67. Mercer, J.: Functions of positive and negative type and their connection with the theory of integral equations. Philos. Trans. R. Soc. Lond. A **209**, 415–446 (1909)
68. Moghaddam, B., Pentland, A.: Probabilistic visual learning for object representation. IEEE Trans. Pattern Anal. Mach. Intell. **19**(7), 696–710 (1997)
69. Mumford, D., Shah, J.: Optimal approximations by piecewise smooth functions and associated variational problems. Commun. Pure Appl. Math. **42**, 577–685 (1989)
70. Nain, D., Yezzi, A., Turk, G.: Vessel segmentation using a shape driven flow. In: MICCAI, Montréal, pp. 51–59 (2003)
71. Neumaier, A., Schneider, T.: Estimation of parameters and eigenmodes of multivariate autoregressive models. ACM Trans. Math. Softw. **27**(1), 27–57 (2001)
72. Osher, S.J., Sethian, J.A.: Fronts propagation with curvature dependent speed: algorithms based on Hamilton–Jacobi formulations. J. Comput. Phys. **79**, 12–49 (1988)
73. Paragios, N., Deriche, R.: Geodesic active regions and level set methods for supervised texture segmentation. Int. J. Comput. Vis. **46**(3), 223–247 (2002)
74. Parent, P., Zucker, S.W.: Trace inference, curvature consistency, and curve detection. IEEE Trans. Pattern Anal. Mach. Intell. **11**(8), 823–839 (1989)
75. Parzen, E.: On the estimation of a probability density function and the mode. Ann. Math. Stat. **33**, 1065–1076 (1962)
76. Rasmussen, C.-E., Williams, C.K.I.: Gaussian Processes for Machine Learning. MIT, Cambridge (2006)
77. Riklin-Raviv, T., Kiryati, N., Sochen, N.: Unlevel sets: geometry and prior-based segmentation. In: Pajdla, T., Hlavac, V. (eds.) European Conference on Computer Vision. LNCS, vol. 3024, pp. 50–61. Springer, Prague (2004)
78. Rochery, M., Jermyn, I., Zerubia, J.: Higher order active contours. Int. J. Comput. Vis. **69**, 27–42 (2006)
79. Rosenblatt, F.: Remarks on some nonparametric estimates of a density function. Ann. Math. Stat. **27**, 832–837 (1956)
80. Rosenfeld, A., Zucker, S.W., Hummel, R.A.: An application of relaxation labeling to line and curve enhancement. IEEE Trans. Comput. **26**(4), 394–403 (1977)
81. Rousson, M., Brox, T., Deriche, R.: Active unsupervised texture segmentation on a diffusion based feature space. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Madison, pp. 699–704 (2003)
82. Rousson, M., Cremers, D.: Efficient kernel density estimation of shape and intensity priors for level set segmentation. In: MICCAI, Palm Springs, vol. 1, pp. 757–764 (2005)
83. Rousson, M., Paragios, N.: Shape priors for level set representations. In: Heyden, A., et al. (eds.) European Conference on Computer Vision, Copenhagen. LNCS, vol. 2351, pp. 78–92. Springer (2002)
84. Rousson, M., Paragios, N., Deriche, R.: Implicit active shape models for 3D segmentation in MRI imaging. In: MICCAI, Saint-Malo. LNCS, vol. 2217, pp. 209–216. Springer (2004)
85. Schmidt, F.R., Farin, D., Cremers, D.: Fast matching of planar shapes in sub-cubic runtime. In: IEEE International Conference on Computer Vision, Rio de Janeiro (2007)
86. Schoenemann, T., Cremers, D.: Globally optimal image segmentation with an elastic shape prior. In: IEEE International Conference on Computer Vision, Rio de Janeiro (2007)
87. Schoenemann, T., Cremers, D.: Introducing curvature into globally optimal image segmentation: minimum ratio cycles on product graphs. In: IEEE International Conference on Computer Vision, Rio de Janeiro (2007)

88. Schoenemann, T., Cremers, D.: Matching non-rigidly deformable shapes across images: a globally optimal solution. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Anchorage (2008)
89. Schoenemann, T., Cremers, D.: A combinatorial solution for model-based image segmentation and real-time tracking. IEEE Trans. Pattern Anal. Mach. Intell. **32**(7), 1153–1164 (2010)
90. Schoenemann, T., Kahl, F., Cremers, D.: Curvature regularity for region-based image segmentation and inpainting: a linear programming relaxation. In: IEEE International Conference on Computer Vision, Kyoto (2009)
91. Schoenemann, T., Schmidt, F.R., Cremers, D.: Image segmentation with elastic shape priors via global geodesics in product spaces. In: British Machine Vision Conference, Leeds (2008)
92. Schwarz, G.: Estimating the dimension of a model. Ann. Stat. **6**, 461–464 (1978)
93. Sebastian, T., Klein, P., Kimia, B.: On aligning curves. IEEE Trans. Pattern Anal. Mach. Intell. **25**(1), 116–125 (2003)
94. Serra, J.: Image Analysis and Mathematical Morophology. Academic, London (1982)
95. Silverman, B.W.: Choosing the window width when estimating a density. Biometrika **65**, 1–11 (1978)
96. Silverman, B.W.: Density Estimation for Statistics and Data Analysis. Chapman and Hall, London (1992)
97. Sundaramoorthi, G., Yezzi, A., Mennucci, A., Sapiro, G.: New possibilities with sobolev active contours. Int. J. Comput. Vis. **84**(2), 113–129 (2009)
98. Sussman, M., Smereka, P., Osher, S.J.: A level set approach for computing solutions to incompressible twophase flow. J. Comput. Phys. **94**, 146–159 (1994)
99. Tsai, A., Wells, W., Warfield, S.K., Willsky, A.: Level set methods in an EM framework for shape classification and estimation. In: MICCAI, Saint-Malo (2004)
100. Tsai, A., Yezzi, A., Wells, W., Tempany, C., Tucker, D., Fan, A., Grimson, E., Willsky, A.: Model–based curve evolution technique for image segmentation. In: IEEE Conference on Computer Vision Pattern Recognition, Kauai, pp. 463–468 (2001)
101. Tsai, A., Yezzi, A.J., Willsky, A.S.: Curve evolution implementation of the Mumford–Shah functional for image segmentation, denoising, interpolation, and magnification. IEEE Trans. Image Process. **10**(8), 1169–1186 (2001)
102. Unal, G., Krim, H., Yezzi, A.Y.: Information-theoretic active polygons for unsupervised texture segmentation. Int. J. Comput. Vis. **62**(3), 199–220 (2005)
103. Unger, M., Pock, T., Cremers, D., Bischof, H.: TVSeg – interactive total variation based image segmentation. In: British Machine Vision Conference (BMVC), Leeds (2008)
104. Wagner, T.J.: Nonparametric estimates of probability densities. IEEE Trans. Inf. Theory **21**, 438–440 (1975)
105. Zhu, S.C., Yuille, A.: Region competition: unifying snakes, region growing, and Bayes/MDL for multiband image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **18**(9), 884–900 (1996)

Optical Flow

Florian Becker, Stefania Petra, and Christoph Schnörr

Contents

1 Introduction 1946
Motivation, Overview 1946
Organization 1947
2 Basic Aspects 1947
Invariance, Correspondence Problem 1947
Assignment Approach, Differential Motion Approach 1950
Two-View Geometry, Assignment and Motion Fields 1956
Early Pioneering Work 1961
Benchmarks 1962
3 The Variational Approach to Optical Flow Estimation 1963
Differential Constraint Equations, Aperture Problem 1963
The Approach of Horn and Schunck 1964
Data Terms 1968
Regularization 1970
Further Extensions 1973
Algorithms 1977
4 The Assignment Approach to Optical Flow Estimation 1981
Local Approaches 1982
Assignment by Displacement Labeling 1984
Variational Image Registration 1987
5 Open Problems and Perspectives 1988
Unifying Aspects: Assignment by Optimal Transport 1988
Motion Segmentation, Compressive Sensing 1992
Probabilistic Modeling and Online Estimation 1994
6 Conclusion 1995

F. Becker (✉)
Heidelberg Collaboratory for Image Processing, University of Heidelberg, Heidelberg, Germany
e-mail: becker@math.uni-heidelberg.de

S. Petra • C. Schnörr
Faculty of Mathematics and Computer Science, Heidelberg University, Heidelberg, Germany
e-mail: petra@math.uni-heidelberg.de; schnoerr@math.uni-heidelberg.de

© Springer Science+Business Media New York 2015

O. Scherzer (ed.), *Handbook of Mathematical Methods in Imaging*,
DOI 10.1007/978-1-4939-0790-8_38

7 Basic Notation 1996
Cross-References 1998
References 1998

Abstract

Motions of physical objects relative to a camera as observer naturally occur in everyday lives and in many scientific applications. Optical flow represents the corresponding motion induced on the image plane. This paper describes the basic problems and concepts related to optical flow estimation together with mathematical models and computational approaches to solve them. Emphasis is placed on common and different modeling aspects and to relevant research directions from a broader perspective. The state of the art and corresponding deficiencies are reported along with directions of future research. The presentation aims at providing an accessible guide for practitioners as well as stimulating research work in relevant fields of mathematics and computer vision.

1 Introduction

Motivation, Overview

Motion of image data belongs to the crucial features that enable low-level image analysis in natural vision systems and in machine vision systems and the analysis of a major part of stored image data in the format of videos, as documented, for instance, by the fast increasing download rate of YouTube. Accordingly, *image motion analysis* has played a key role from the beginning of research in mathematical and computational approaches to image analysis.

Figure 1 illustrates few application areas of image processing, among many others, where image motion analysis is deeply involved. Mathematical models for analyzing such image sequences boil down to models of a specific instance of the general data analysis task, that is, to fuse prior knowledge with information given by observed image data. While adequate prior knowledge essentially depends on the application area as Fig. 1 indicates, the processing of observed data mainly involves basic principles that apply to *any* image sequence. Correspondingly, the notion of *optical flow*, informally defined as *determining the apparent instantaneous velocity of image structure*, emphasizes the application-independent aspects of this basic image analysis task.

Due to this independency, optical flow algorithms provide a key component for numerous approaches to *applications across different fields*. Major examples include motion compensation for video compression, structure from motion to estimate 3-D scene layouts from image sequences, visual odometry, and incremental construction of mappings of the environment by autonomous systems, estimating vascular wall shear stress from blood flow image sequences for biomedical diagnosis, to name just a few.

This chapter aims at providing a concise and up-to-date account of mathematical models of optical flow estimation. Basic principles are presented along with various

Fig. 1 Some application areas of image processing that essentially rely on image motion analysis. *Left*: scene analysis (depth, independently moving objects) with a camera mounted in a car. *Center*: flow analysis in remote sensing. *Right*: measuring turbulent flows by particle image velocimetry

prior models. Application-specific aspects are only taken into account at a general level of mathematical modeling (e.g., geometric or physical prior knowledge). Model properties favoring a particular direction of modeling are highlighted, while keeping an eye on common aspects and open problems. Conforming to the editor's guidelines, references to the literature are confined to a – subjectively defined – essential minimum.

Organization

Section 2 introduces a dichotomy of models used to present both essential differences and common aspects. These classes of models are presented in Sects. 3 and 4. The former class comprises those algorithms that perform best on current benchmark datasets. The latter class becomes increasingly more important in connection with motion analysis of novel, challenging classes of image sequences and videos. While both classes merely provide different viewpoints on the same subject – optical flow estimation and image motion analysis – distinguishing them facilitates the presentation of various facets of relevant mathematical models in current research. Further relationships, unifying aspects together with some major open problems and research directions, are addressed in Sect. 5.

2 Basic Aspects

Invariance, Correspondence Problem

Image motion computation amounts to define some notion of *invariance* and the recognition in subsequent time frames of *corresponding objects*, defined by local prominent image structure in terms of a feature mapping $g(x)$ whose values are assumed to be conserved during motion. As Fig. 2, left panel, illustrates, invariance only holds approximately due to the imaging process and changes of viewpoint and

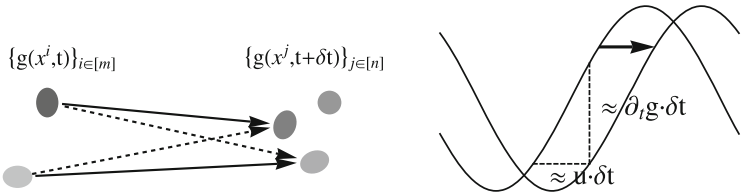


Fig. 2 *Left*: image motion can only be computed by recognizing objects as the same in subsequent time frames, based on some notion of equivalence (invariance) and some distance function. In low-level vision, "object" means some prominent local image structure in terms a feature mapping $g(x,t) \in \mathbb{R}^p$, $p \geq 1$. The correspondence problem amounts to compute a corresponding assignment $\{g(x^i,t)\}_{i\in[m]} \to \{g(x^j,t+\delta t)\}_{j\in[n]}$. The corresponding objective defines the data term of a variational approach. *Right*: differential approaches to image motion computation are based on smooth feature mappings $g(x,t)$ and aim at solving the assignment problem $g\big(x(t),t\big) \to g\big(x(t+\delta t),t+\delta t\big)$. The figure illustrates the basic case of a scalar-valued signal $g(x,t)$ translating with constant speed u and the estimate (14) based on the differential motion approach, as discussed in section "Assignment Approach, Differential Motion Approach"

illumination. Consequently, some *distance function*

$$\rho\big(g(x^j,t+\delta t) - g(x^i,t)\big) \tag{1}$$

has to be used in order to compute an *optimal assignment*

$$\{g(x^i,t)\}_{i\in[m]} \to \{g(x^j,t+\delta t)\}_{j\in[n]}. \tag{2}$$

A vast literature exists on definitions of feature mappings $g(x,t)$, distance functions, and their empirical evaluation in connection with image motion. Possible definitions include

- Image gray value or color,
- Gray value or color gradient,
- Output of analytic band-pass filters (e.g., [1,2]),
- More complex feature descriptors including SIFT [3] and SURF [4],
- Censor voting, [5], local patches or feature groupings,

together with a corresponding invariance assumption, i.e., that $g(x,t)$ is conserved during motion (cf. Fig. 2, left panel). Figure 3 illustrates the most basic approaches used in the literature. Recent examples adopting a more geometric viewpoint on feature descriptors and studying statistical principles of patch similarity include [6,7].

For further reference, some basic distance functions $\rho\colon \mathbb{R}^p \to \mathbb{R}_+$ are introduced below that are commonly applied in connection with feature mappings $g(x)$ and partly parametrized by $\lambda > 0$ and $0 < \varepsilon \ll 1$. For closely related functions and the nomenclature in computer vision, see, e.g., [8].

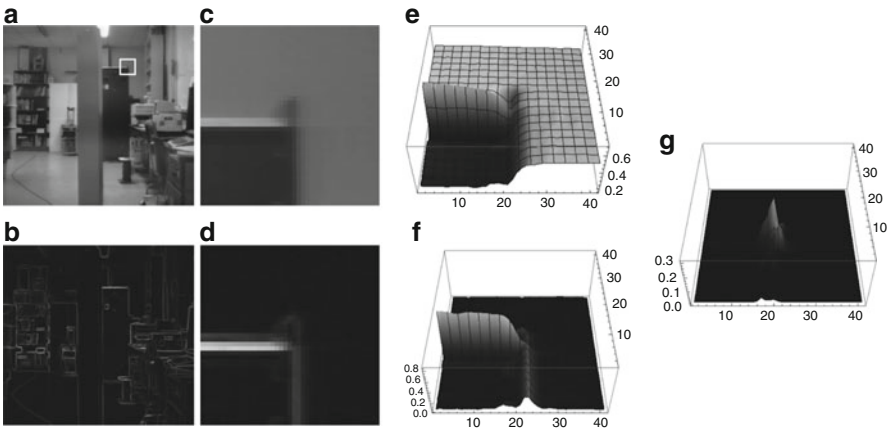


Fig. 3 (**a**) Lab scene (©CMU image database) and (**b**) gradient magnitude that provides the basis for a range of feature mappings $g(x,t)$. The image section indicated in (**a**) is shown in (**c**), and (**d**) shows the same section extracted from (**b**). Panels (**e**) and (**f**) illustrate these sections as surface plots. Panel (**g**) shows a feature map responding to crossing gray-value edges. (**c**), (**d**), and (**g**) correspond to the most basic examples of feature mappings $g(x,t)$ used in the literature to compute image motion, based on a corresponding invariance assumption (cf. Fig. 2, left panel) that is plausible for video frame rates

$$\rho_2^2(z) := \|z\|^2 \qquad \text{squared } \ell^2 \text{ distance,} \tag{3a}$$

$$\rho_2(z) := \|z\| \qquad \ell^2 \text{ distance,} \tag{3b}$$

$$\rho_{2,\varepsilon}(z) := \sqrt{\|z\|^2 + \varepsilon^2} - \varepsilon \qquad \text{smoothed } \ell^2 \text{ distance,} \tag{3c}$$

$$\rho_1(z) := \|z\|_1 \qquad \ell^1 \text{ distance,} \tag{3d}$$

$$\rho_{1,\varepsilon}(z) := \sum_{i\in[p]} \rho_{2,\varepsilon}(z_i) \qquad \text{smoothed } \ell^1 \text{ distance,} \tag{3e}$$

$$\rho_{2,\lambda}(z) := \min\{\|z\|^2, \lambda^2\} \qquad \text{truncated squared } \ell^2 \text{ distance,} \tag{3f}$$

$$\rho_{2,\lambda,\varepsilon}(z) := -\varepsilon \log\left(\mathrm{e}^{-\|z\|^2/\varepsilon} + \mathrm{e}^{-\lambda^2/\varepsilon}\right) \qquad \text{smoothed tr. sq. } \ell^2 \text{ distance.} \tag{3g}$$

Figure 4 illustrates these convex and non-convex distance functions. Functions $\rho_{1,\varepsilon}$ and $\rho_{2,\varepsilon}$ constitute specific instances of the general smoothing principle to replace a lower semicontinuous, positively homogeneous, and sublinear function $\rho(z)$ by a smooth proper convex function $\rho_\varepsilon(z)$, with $\lim_{\varepsilon \searrow 0} \varepsilon\rho_\varepsilon(z/\varepsilon) = \rho(z)$ (cf., e.g., [9]). Function $\rho_{2,\lambda,\varepsilon}(z)$ utilizes the log-exponential function [10, Ex. 1.30] to uniformly approximate $\rho_{2,\lambda}$ as $\varepsilon \searrow 0$.

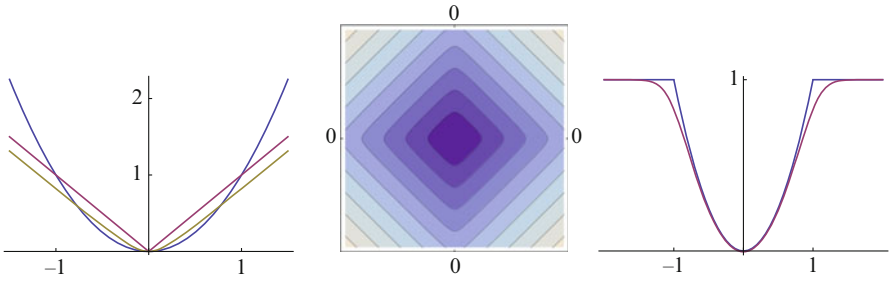


Fig. 4 *Left*: convex distance functions (3a)–(3c). *Center*: level lines of the distance function $\rho_{1,\varepsilon}$ (3e). *Right*: non-convex distance functions (3f) and (3g)

Assignment Approach, Differential Motion Approach

Definitions

Two basic approaches to image motion computation can be distinguished.

Assignment Approach, Assignment Field This approach aims to determine an *assignment* of finite sets of spatially discrete features in subsequent frames of a given image sequence (Fig. 2, left panel). The vector field

$$u(x,t), \qquad x^j = x^i + u(x^i,t), \tag{4}$$

representing the assignment in Eq. (2), is called *assignment field*. This approach conforms to the basic fact that image sequences $f(x,t),\ (x,t) \in \Omega \times [0,T]$ are recorded by *sampling* frames

$$\{f(x,k\cdot\delta t)\}_{k\in\mathbb{N}} \tag{5}$$

along the time axis.

Assignment approaches to image motion will be considered in Sect. 4.

Differential Motion Approach, Optical Flow Starting point of this approach is the *invariance assumption* (section "Invariance, Correspondence Problem") that observed values of some feature map $g(x,t)$ are conserved during motion,

$$\frac{\mathrm{d}}{\mathrm{d}t} g\big(x(t),t\big) = 0. \tag{6}$$

Evaluating this condition yields information about the trajectory $x(t)$ that represents the motion path of a particular feature value $g\big(x(t)\big)$. The corresponding vector field

$$\dot{x}(t) = \frac{\mathrm{d}}{\mathrm{d}t} x(t), \qquad x \in \Omega \tag{7}$$

is called *motion field* whose geometric origin will be described in section "Two-View Geometry, Assignment and Motion Fields." *Estimates*

$$u(x,t) \approx \dot{x}(t), \qquad x \in \Omega \tag{8}$$

of the motion field based on some observed time-dependent feature map $g(x,t)$ are called *optical flow fields*.
Differential motion approaches will be considered in Sect. 3.

Common Aspects and Differences

The assignment approach and the differential approach to image motion are closely related. In fact, for small temporal sampling intervals,

$$0 < \delta t \ll 1, \tag{9}$$

one may expect that the optical flow field multiplied by δt, $u(x,t)\cdot\delta t$, closely approximates the corresponding assignment field. The *same* symbol u is therefore used in (4) and (8) to denote the respective vector fields.

A conceptual difference between both approaches is that the ansatz (6) entails the assumption of a spatially differentiable feature mapping $g(x,t)$, whereas the assignment approach requires prior decisions done at a preprocessing stage that localize the feature sets (2) to be assigned. The need for additional processing in the latter case contrasts with the limited applicability of the differential approach: *The highest spatial frequency limits the speed of image motion $\|u\|$ that can be estimated reliably*:

$$\max\left\{\|\omega_x\|_\infty, \|u(x)\|\|\omega_x\| : \omega_x \in \mathrm{supp}\hat{g}(\omega),\ x \in \Omega\right\} \leq \frac{\pi}{6}. \tag{10}$$

The subsequent section details this bound in the most simple setting for a specific but common filter choice for estimating partial derivatives $\partial_i g$.

Differential Motion Estimation: Case Study (1D)

Consider a scalar signal $g(x,t) = f(x,t)$ moving at constant speed (cf. Fig. 2, right panel),

$$\dot{x}(t) = \dot{x} = u, \qquad g\big(x(t),t\big) = g\big(x(0)+ut,t\big). \tag{11}$$

Note that the two-dimensional function $g(x,t)$ is a very special one generated by motion. Using the shorthands

$$x := x(0), \quad g_0(x) := g(x,0), \tag{12}$$

$g(x,t)$ corresponds to the translated one-dimensional signal

$$g(x,t) = g_0(x - ut) \tag{13}$$

due to the assumption $g\big(x(t),t\big) = g\big(x(0),0\big) = g_0(x)$.

Evaluating (6) at $t = 0$, $x = x(0)$ yields

$$u = -\frac{\partial_t g_0(x)}{\partial_x g_0(x)} \quad \text{if} \quad \partial_x g_0(x) \neq 0. \tag{14}$$

Application and validity of this equation in practice depends on two further aspects: Only sampled values of $g(x,t)$ are given, and the right-hand side has to be computed numerically. Both aspects are discussed next in turn.

1. In practice, samples are observed

$$\{g(k \cdot \delta x, t\delta t)\}_{k,t\in\mathbb{N}} = \{g(k,t)\}_{k,t\in\mathbb{N}}, \qquad \delta x = \delta t = 1, \tag{15}$$

with the sampling interval scaled to 1 without loss of generality. The Nyquist-Shannon sampling theorem imposes the constraint

$$\text{supp}|\hat{g}(\omega)| \subset [0,\pi)^2, \qquad \omega = (\omega_x, \omega_t)^\top \tag{16}$$

where

$$\hat{g}(\omega) = \mathcal{F}g(\omega) = \int_{\mathbb{R}^2} g(x,t) e^{-i\langle\omega,\binom{x}{t}\rangle} dx dt \tag{17}$$

denotes the Fourier transform of $g(x,t)$. Trusting in the sensor, it may be savely assumed that $\text{supp}|\hat{g}_0(\omega_x)| \subset [0,\pi)$. But what about the second coordinate t generated by motion? Does it obey (16) such that the observed samples (15) truly represent the one-dimensional video signal $g(x,t)$?

To answer this question, consider the specific case $g_0(x) = \sin(\omega_x x)$, $\omega_x \in [0,\pi]$ – see Fig. 5. Equation (13) yields $g(x,t) = \sin\big(\omega_x(x - ut)\big)$. Condition (15) then requires that, for *every* location x, the one-dimensional time signal $g_x(t) := g(x,t)$ satisfies $\text{supp}|\hat{g}_x(\omega_t)| \subset [0,\pi)$. Applying this to the example yields

$$g_x(t) = \sin(\omega_t t + \varphi_0), \qquad \omega_t := -\omega_x u, \quad \varphi_0 := \omega_x x, \tag{18}$$

and hence the condition

$$|\omega_t| \in [0,\pi) \qquad \Leftrightarrow \qquad |u| < \frac{\pi}{\omega_x}. \tag{19}$$

It implies that Eq. (14) is only valid if, depending on the spatial frequency ω_x, the velocity u is sufficiently small.

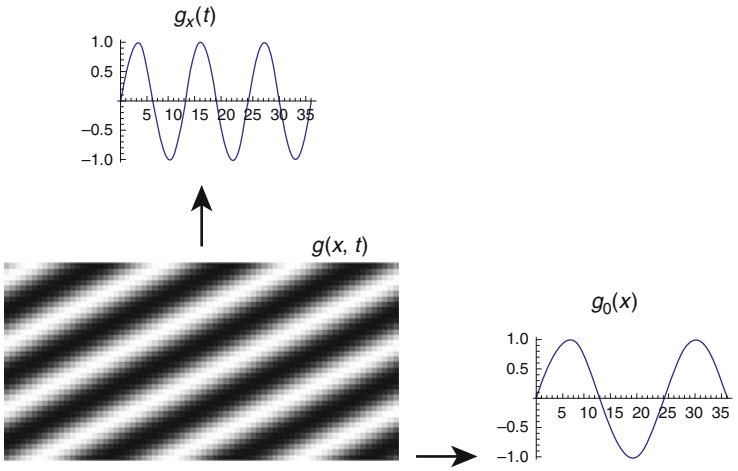


Fig. 5 A sinusoid $g_0(x)$ with angular frequency $\omega_x = \pi/12$, translating with velocity $u = 2$, generates the function $g(x,t)$. The angular frequency of the signal $g_x(t)$ observed at a fixed position x equals $|\omega_t| = u \cdot \omega_x = \pi/6$ due to (18). It meets the upper bound further discussed in connection with Fig. 6 that enables accurate numerical computation of the partial derivatives of $g(x,t)$

This reasoning and the conclusion apply to general functions $g(x,t)$, $x \in \mathbb{R}^d$ in the form of (10), which additionally takes into account the effect of derivative estimation, discussed next.

2. Condition (19) has to be further restricted in practice, depending on how the partial derivatives of the r.h.s. of Eq. (14) are *numerically* computed using the observed samples (15). The Fourier transform

$$\mathcal{F}\big(\partial^\alpha g\big)(\omega) = i^{|\alpha|}\omega^\alpha \hat{g}(\omega), \qquad \omega \in \mathbb{R}^{d+1} \tag{20}$$

generally shows that taking partial derivatives of order $|\alpha|$ of $g(x,t)$, $x \in \mathbb{R}^d$, corresponds to high-pass filtering that amplifies noise. If $g(x,t)$ is vector valued, then the present discussion applies to the computation of partial derivatives $\partial^\alpha g_i$ of any component $g_i(x,t)$, $i \in [p]$.

To limit the influence of noise, partial derivatives of the *low-pass filtered* feature mapping g are computed. This removes noise and smoothes the signal, and subsequent computation of partial derivatives becomes more accurate. Writing $g(x)$, $x \in \mathbb{R}^{d+1}$, instead of $g(x,t)$, $x \in \mathbb{R}^d$, to simplify the following formulas, low-pass filtering of g with the impulse response $h(x)$ means the convolution

$$g_h(x) := (h * g)(x) = \int_{\mathbb{R}^{d+1}} h(x-y)g(y)\mathrm{d}y, \qquad \hat{g}_h(\omega) = \hat{h}(\omega)\,\hat{g}(\omega) \tag{21}$$

whose Fourier transform corresponds to the multiplication of the respective Fourier transforms. Applying (20) yields

$$\mathcal{F}\big(\partial^\alpha g_h\big)(\omega) = i^{|\alpha|}\omega^\alpha\big(\hat{h}(\omega)\,\hat{g}(\omega)\big) = \big(i^{|\alpha|}\omega^\alpha\hat{h}(\omega)\big)\hat{g}(\omega). \tag{22}$$

Thus, computing the partial derivative of the filtered function g_h can be computed by convolving g with the partial derivative of the impulse response $\partial^\alpha h$. As a result, the approximation of the partial derivative of g reads

$$\partial^\alpha g(x) \approx \partial^\alpha g_h(x) = \big((\partial^\alpha h) * g\big)(x). \tag{23}$$

The most common choice of h is the isotropic Gaussian low-pass filter

$$h_\sigma(x) := \frac{1}{(2\pi\sigma^2)^{d/2}}\exp\Big(-\frac{\|x\|^2}{2\sigma^2}\Big) = \prod_{i\in[d]} h_\sigma(x_i), \qquad \sigma > 0 \tag{24}$$

that factorizes (called *separable filter*) and therefore can be implemented efficiently. The corresponding filters $\partial^\alpha h_\sigma(x)$, $|\alpha| \geq 1$ are called *Derivative-of-Gaussian (DoG) filters*.

To examine its effect, it suffices to consider any coordinate due to factorization, that is, the one-dimensional case. Figure 6 illustrates that values $\sigma \geq 1$ lead to filters that are sufficiently band limited so as to conform to the sampling theorem. The price to pay for effective noise suppression however is a more restricted range $\text{supp}|\mathcal{F}\big(g(x,t)\big)| = [0, \omega_{x,\max}]$, $\omega_{x,\max} \ll \pi$ that observed image sequence functions have to satisfy, so as to enable accurate computation of partial derivatives and in turn accurate motion estimates based on the differential approach. Figure 5 further details and illustrates this crucial fact.

Assignment or Differential Approach?

For image sequence functions $g(x,t)$ satisfying the assumptions necessary to evaluate the key equation (6), the differential motion approach is more convenient. Accordingly, much work has been devoted to this line of research up to now. In particular, sophisticated multiscale representations of $g(x,t)$ enable to estimate larger velocities of image motion using smoothed feature mapping g (cf. section "Multiscale"). As a consequence, differential approaches rank top at corresponding benchmark evaluations conforming to the underlying assumptions [11], and efficient implementations are feasible [12, 13].

On the other hand, the inherent limitations of the differential approach discussed above become increasingly more important in current applications, like optical flow computation for traffic scenes taken from a moving car at high speed. Figure 1, right panel, shows another challenging scenario where the spectral properties $\hat{g}(\omega_x, \omega_t)$ of the image sequence function and the velocity fields to be estimated render

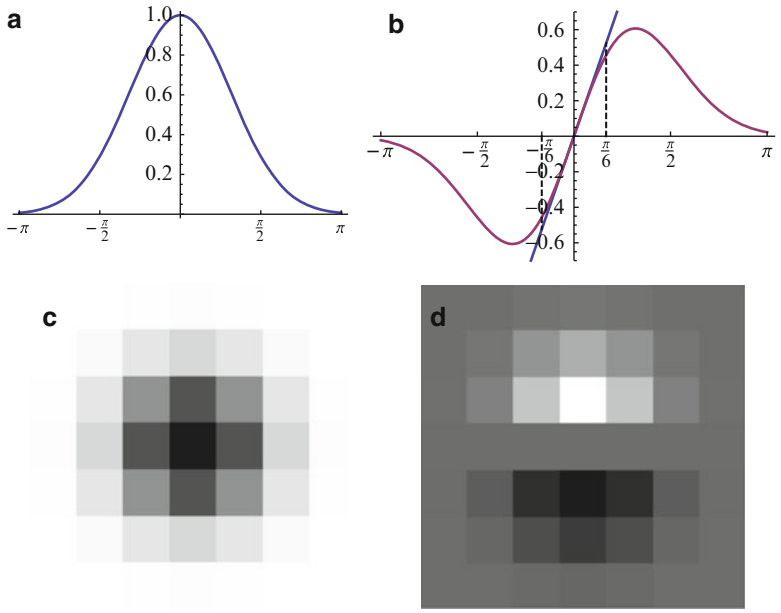


Fig. 6 **(a)** Fourier transform $\hat{h}_\sigma(\omega)$ of the Gaussian low pass (24), $\sigma = 1$. For values $\sigma \geq 1$, it satisfies the sampling condition $\mathrm{supp}|\hat{h}_\sigma(\omega)| \subset [0, \pi)$ sufficiently accurate. **(b)** The Fourier transform of the *Derivative-of-Gaussian (DoG)* filter $\frac{\mathrm{d}}{\mathrm{d}x} h_\sigma(x)$ illustrates that for $|\omega| \leq \pi/6$ (partial) derivatives are accurately computed, while noise is suppressed at higher angular frequencies. **(c)**, **(d)** The impulse responses $h_\sigma(x,t)$ and $\partial_t h_\sigma(x,t)$ up to size $|x|, |t| \leq 2$. Application of the latter filter together with $\partial_x h_\sigma(x,t)$ to the function $g(x,t)$ discussed in connection with Fig. 5 and evaluation of Eq. (14) yield the estimate $u = 2.02469$ at all locations (x,t) where $\partial_x g(x,t) \neq 0$

application of the differential approach difficult, if not impossible. In such cases, the assignment approach is the method of choice.

Combining both approaches in a complementary way seems most promising: Robust assignments enable to cope with fast image motions and a differential approach turns these estimates into spatially dense vector fields. This point is taken up in section "Unifying Aspects: Assignment by Optimal Transport."

Basic Difficulties of Motion Estimation

This section concludes with a list of some basic aspects to be addressed by any approach to image motion computation:

(i) Definition of a feature mapping g assumed to be conserved during motion (section "Invariance, Correspondence Problem").
(ii) Coping with lack of invariance of g, change of appearance due to varying viewpoint and illumination (sections "Handling Violation of the Constancy Assumption" and "Patch Features").
(iii) Spatial sparsity of distinctive features (section "Regularization").

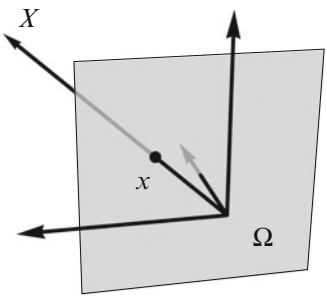


Fig. 7 The basic pinhole model of the mathematically ideal camera. Scene points X are mapped to image points x by perspective projection

(iv) Coping with ambiguity of locally optimal feature matches (section "Assignment by Displacement Labeling").
(v) Occlusion and disocclusion of features.
(vi) Consistent integration of available prior knowledge, regularization of motion field estimation (sections "Geometrical Prior Knowledge" and "Physical Prior Knowledge").
(vii) Runtime requirements (section "Algorithms").

Two-View Geometry, Assignment and Motion Fields

This section collects few basic relationships related to the Euclidean motion of a perspective camera relative to a 3-D scene that induces both the assignment field and the motion field on the image plane, as defined in section "Definitions" by (4) and (7). Figures 7 and 12 illustrate these relationships. References [14, 15] provide comprehensive expositions.

It is pointed out once more that assignment and motion fields are purely *geometrical* concepts. The explicit expressions (43) and (53b) illustrate how discontinuities of these fields correspond to *discontinuities of depth* or to *motion boundaries* that separate regions in the image plane of scene objects (or the background) with different motions relative to the observing camera. Estimates of either field will be called *optical flow*, to be discussed in subsequent sections.

Two-View Geometry

Scene and corresponding image points are denoted by $X \in \mathbb{R}^3$ and $x \in \mathbb{R}^2$, respectively. Both are incident with the line λx, $\lambda \in \mathbb{R}$, through the origin. Such lines are points of the projective plane denoted by $y \in \mathbb{P}^2$. The components of y are called *homogeneous coordinates* of the image point x, whereas x and X are the *inhomogeneous coordinates* of image and scene points, respectively. Note that y stands for *any* representative point on the ray connecting x and X. In other words, when using homogeneous coordinates, scale factors do not matter. This equivalence is denoted by

$$y \simeq y' \quad \Leftrightarrow \quad y = \lambda y', \quad \lambda \neq 0. \tag{25}$$

Figure 7 depicts the mathematical model of a pinhole camera with the image plane located at $X_3 = 1$. Perspective projection corresponding to this model connects homogeneous and inhomogeneous coordinates by

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \frac{1}{y_3} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}. \tag{26}$$

A *particular representative* y with *unknown depth* $y_3 = X_3$ equals the scene point X. This reflects the fact that scale cannot be inferred from a single image. The 3-D space $\mathbb{R}^3 \setminus \{0\}$ corresponds to the affine chart $\{y \in \mathbb{P}^2 \colon y_3 \neq 0\}$ of the manifold $\mathbb{P}^2$.

Similar to representing an image point x by homogeneous coordinates y, it is common to represent scene points $X \in \mathbb{R}^3$ by *homogeneous coordinates* $Y = (Y_1, Y_2, Y_3, Y_4)^\top \in \mathbb{P}^3$, in order to linearize transformations of 3-D space. The connection analogous to (26) is

$$X = \frac{1}{Y_4} \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix}. \tag{27}$$

Rigid (Euclidean) transformations are denoted by $\{h, R\} \in \mathsf{SE}(3)$ with translation vector h and proper rotation matrix $R \in \mathsf{SO}(3)$ characterized by $R^\top R = I$, $\det R = +1$. Application of the transformation to a scene point X and some representative Y reads

$$RX + h \qquad \text{and} \qquad QY, \qquad Q := \begin{pmatrix} R & h \\ 0^\top & 1 \end{pmatrix}, \tag{28}$$

whereas the inverse transformation $\{-R^\top h, R^\top\}$ yields

$$R^\top (X - h) \qquad \text{and} \qquad Q^{-1} Y, \qquad Q^{-1} = \begin{pmatrix} R^\top & -R^\top h \\ 0^\top & 1 \end{pmatrix}. \tag{29}$$

The nonlinear operation (26), entirely rewritten with homogeneous coordinates, takes the linear form

$$y = PY, \qquad P = \begin{pmatrix} 1\,0\,0\,0 \\ 0\,1\,0\,0 \\ 0\,0\,1\,0 \end{pmatrix} = (I_{3\times 3}, 0), \tag{30}$$

with the *projection matrix* P and *external* or *motion parameters* $\{h, R\}$. In practice, additional *internal parameters* characterizing real cameras to the first order of

approximation are taken into account in terms of a *camera matrix* K and the corresponding modification of (30),

$$y = PY, \qquad P = K(I_{3\times 3}, 0). \tag{31}$$

As a consequence, the transition to *normalized (calibrated)* coordinates

$$\tilde{y} := K^{-1} y \tag{32}$$

corresponds to an *affine* transformation of the image plane.

Given an image point x, taken with a camera in the canonical position (30), the corresponding ray meets the scene point X; see Figure 12b. This ray projects in a second image, taken with a second camera positioned by $\{h, R\}$ relative to the first camera and with projection matrix

$$P' = K' R^{\top}(I, -h), \tag{33}$$

to the line l', on which the projection x' of X *corresponding* to x must lie. Turning to homogeneous coordinates, an elementary computation shows that the *fundamental matrix*

$$F := K'^{-\top} R^{\top} [h]_{\times} K^{-1} \tag{34}$$

maps y to the *epipolar line* l',

$$l' = Fy. \tag{35}$$

This relation is symmetrical in that $F^{\top}$ maps y' to the corresponding epipolar line l in the first image,

$$l = F^{\top} y'. \tag{36}$$

The *epipoles* e, e' are the image points corresponding to the projection centers. Because they lie on l and l' for *any* x' and x, respectively, it follows that

$$Fe = 0, \qquad F^{\top} e' = 0. \tag{37}$$

The incidence relation $x' \in l'$ algebraically reads $\langle l', y' \rangle = 0$. Hence by (35),

$$\langle y', Fy \rangle = 0 \tag{38}$$

This key relation constrains the correspondence problem $x \leftrightarrow x'$ for arbitrary two views of the same unknown scene point X. Rewriting (38) in terms of *normalized* coordinates by means of (32) yields

$$\langle y', Fy\rangle = \langle K'^{-1}y', K'^{\top}FK(K^{-1}y)\rangle = \langle K'^{-1}y', E(K^{-1}y)\rangle \tag{39}$$

with the *essential matrix* E that, due to (34) and the relation $[Kh]_\times \simeq K^{-\top}[h]_\times K^{-1}$, is given by

$$E = K'^{\top}FK = R^{\top}[h]_\times. \tag{40}$$

Thus, essential matrices are parametrized by transformations $\{h, R\} \in \mathsf{SE}(3)$ and therefore form a smooth manifold embedded in $\mathbb{R}^{3\times 3}$.

Assignment Fields

Throughout this section, the internal camera parameters K are assumed to be known, and hence normalized coordinates (32) are used. As a consequence,

$$K = I \tag{41}$$

is set in what follows.

Suppose some motion h, R of a camera relative to a 3-D scene causes the image point x of a fixed scene point X to move to x' in the image plane. The corresponding *assignment vector* $u(x)$ represents the *displacement* of x in the image plane,

$$x' = x + u(x), \tag{42}$$

which due to (29) and (26) is given by

$$u(x) = \frac{1}{\langle r^3, X - h\rangle}\begin{pmatrix}\langle r^1, X - h\rangle \\ \langle r^2, X - h\rangle\end{pmatrix} - \frac{1}{X_3}\begin{pmatrix}X_1 \\ X_2\end{pmatrix}. \tag{43}$$

Consider the special case of pure translation, i.e., $R = I$, $r^i = e^i$, $i = 1, 2, 3$. Then

$$u(x) = \frac{1}{X_3 - h_3}\begin{pmatrix}X_1 - h_1 \\ X_2 - h_2\end{pmatrix} - \frac{1}{X_3}\begin{pmatrix}X_1 \\ X_2\end{pmatrix} \tag{44a}$$

$$= \frac{1}{\tilde{h}_3 - 1}\left(\begin{pmatrix}\tilde{h}_1 \\ \tilde{h}_2\end{pmatrix} - \tilde{h}_3\begin{pmatrix}x_1 \\ x_2\end{pmatrix}\right), \qquad \tilde{h} := \frac{1}{X_3}h. \tag{44b}$$

The image point x_e where the vector field u vanishes, $u(x_e) = 0$, is called *focus of expansion (FOE)*

$$x_e = \frac{1}{\tilde{h}_3}\begin{pmatrix}\tilde{h}_1 \\ \tilde{h}_2\end{pmatrix}. \tag{45}$$

x_e corresponds to the epipole $y = e$ since $Fe \simeq R^{\top}[h]_\times h = 0$.

Next the transformation is computed of the image plane induced by the motion of the camera in terms of projection matrices $P = (I, 0)$ and $P' = R^\top(I, -h)$ relative to a plane in 3-D space

$$\langle n, X\rangle - d = n_1X_1 + n_2X_2 + n_3X_3 - d = 0, \tag{46}$$

with unit normal n, $\|n\| = 1$ and with signed distance d of the plane from the origin 0. Setting $p = \binom{n}{-d}$, Eq. (46) reads

$$\langle p, Y\rangle = 0. \tag{47}$$

In order to compute the point X on the plane satisfying (46) that projects to the image point y, the ray $Y(\lambda) = \binom{\lambda y}{1}$, $\lambda \in \mathbb{R}$, is intersected with the plane.

$$\langle p, Y(\lambda)\rangle = \lambda\langle n, y\rangle - d = 0 \;\Rightarrow\; \lambda = \frac{d}{\langle n, y\rangle},\; Y = \begin{pmatrix} \frac{d}{\langle n,y\rangle}y \\ 1 \end{pmatrix} \simeq \begin{pmatrix} y \\ \frac{1}{d}\langle n, y\rangle \end{pmatrix}. \tag{48}$$

Projecting this point onto the second image plane yields

$$\begin{aligned} y' &= P'Y(\lambda) = R^\top\left(y - \frac{1}{d}\langle n, y\rangle h\right) = R^\top(I - \frac{1}{d}hn^\top)y \\ &=: Hy \end{aligned} \tag{49}$$

Thus, moving a camera relative to a 3-D plane induces a *homography* (projective transformation) H of $\mathbb{P}^2$ which by virtue of (26) yields an assignment field $u(x)$ with *rational* components.

Motion Fields

Motion fields (7) are the instantaneous (differential) version of assignment fields. Consider a smooth path $\{h(t), R(t)\} \subset \mathsf{SE}(3)$ through the identity $\{0, I\}$ and the corresponding path of a scene point $X \in \mathbb{R}^3$

$$X(t) = h(t) + R(t)X, \qquad X = X(0). \tag{50}$$

Let $R(t)$ be given by a rotational axis $q \in \mathbb{R}^3$ and a rotation angle $\varphi(t)$. Using Rodrigues' formula and the skew-symmetric matrix $[q]_\times \in \mathfrak{so}(3)$ with $\dot\varphi = \dot\varphi(0) := \|q\|$, matrix $R(t)$ takes the form

$$R(t) = \exp(t[q]_\times) = I + \frac{\sin(\dot\varphi t)}{\dot\varphi t}t[q]_\times + \frac{1 - \cos(\dot\varphi t)}{(\dot\varphi t)^2}t^2[q]_\times^2. \tag{51}$$

Equation (50) then yields

$$\dot X(0) = v + [q]_\times X, \qquad v := \dot h(0), \tag{52}$$

where v is the translational velocity at $t = 0$. Differentiating (26) with $y = X$ (recall assumption (41)) and inserting (52) give

$$\frac{\mathrm{d}}{\mathrm{d}t}x = \frac{1}{X_3^2}\begin{pmatrix} X_3\dot{X}_1 - X_1\dot{X}_3 \\ X_3\dot{X}_2 - X_2\dot{X}_3 \end{pmatrix} = \frac{1}{X_3}\begin{pmatrix} \dot{X}_1 - x_1\dot{X}_3 \\ \dot{X}_2 - x_2\dot{X}_3 \end{pmatrix} \tag{53a}$$

$$= \frac{1}{X_3}\left[\begin{pmatrix} v_1 \\ v_2 \end{pmatrix} - v_3\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}\right] + \begin{pmatrix} q_2 - q_3x_2 - q_1x_1x_2 + q_2x_1^2 \\ -q_1 + q_3x_1 + q_2x_1x_2 - q_1x_2^2 \end{pmatrix}. \tag{53b}$$

Comparing (53b) to (43) and (44b) shows a similar structure of the translational part with FOE

$$x_v := \frac{1}{v_3}\begin{pmatrix} v_1 \\ v_2 \end{pmatrix}, \tag{54}$$

whereas the rotational part merely contributes an incomplete second-order degree polynomial to each component of the motion field that does not depend on the scene structure in terms of the depth X_3.

Consider the special case of a motion field induced by the relative motion of a camera to a 3-D plane given by (46) and write

$$\frac{1}{X_3} = \frac{1}{d}\left(n_3 + \begin{pmatrix} n_1 \\ n_2 \end{pmatrix}^\top \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}\right). \tag{55}$$

Insertion into (53b) shows that the overall expression for the motion fields takes a simple polynomial form.

Early Pioneering Work

It deems proper to the authors to refer at least briefly to early pioneering work related to optical flow estimation, as part of a survey paper. The following references constitute just a small sample of the rich literature.

The information of motion fields, induced by the movement of an observer relative to a 3-D scene, was picked out as a central theme more than three decades ago [16, 17]. Kanatani [18] studied the representation of $\mathsf{SO}(3)$ and invariants in connection with the space of motion fields induced by the movement relative to a 3-D plane. Approaches to estimating motion fields followed soon, by determining optical flow from local image structure [19–24]. Poggio and Verri [25] pointed out both the inexpedient, restrictive assumptions making the invariance assumption (6) hold in the simple case $g(x) = f(x)$ (e.g., Lambertian surfaces in the 3-D scene) and the stability of structural (topological) properties of motion fields (like, e.g., the FOE (45)). The local detection of image translation as orientation in spatiotemporal frequency space, based on the

energy and the phase of collections of orientation-selective complex-valued band-pass filters (low-pass filters shifted in Fourier space, like, e.g., Gabor filters), was addressed by [26–28], partially motivated by related research on natural vision systems.

The variational approach to optical flow was pioneered by Horn and Schunck [29], followed by various extensions [30–32] including more mathematically oriented accounts [33–35]. The work [36] classified various convex variational approaches that have unique minimizers.

The computation of discontinuous optical flow fields, in terms of piecewise parametric representations, was considered by [8, 37], whereas the work [38] studied the information contained in correspondences induced by motion fields over a longer time period. Shape-based optimal control of flows determined on discontinuous domains as control variable was introduced in [39], including the application of shape derivative calculus that became popular later on in connection with level sets. Markov random fields and the Bayesian viewpoint on the nonlocal inference of discontinuous optical flow fields were introduced in [40]. The challenging aspects of estimating both motion fields and their segmentation in a spatiotemporal framework, together with inferring the 3-D structure, have remained a topic of research until today.

This brief account shows that most of the important ideas appeared early in the literature. On the other hand, it took many years until first algorithms made their way into industrial applications. A lot of work remains to be done by addressing various basic and applied research aspects. In comparison to the fields of computer vision, computer science, and engineering, not much work has been done by the mathematical community on motion-based image sequence analysis.

Benchmarks

Starting with the first systematic evaluation in 1994 by Baron et al. [41], benchmarks for optical flow methods have stimulated and steered the development of new algorithms in this field. The *Middlebury database* [42] further accelerated this trend by introducing an online ranking system and defining challenging data sets, which specifically address different aspects of flow estimation such as large displacements or occlusion.

The recently introduced *KITTI Vision Benchmark Suite* [43] concentrates on *outdoor* automotive sequences that are affected by disturbances such as illumination changes and reflections, which optical flow approaches are expected to be robust against.

While real imagery requires sophisticated measurement equipment to capture reliable reference information, *synthetic* sequences such as the novel *MPI Sintel Flow Dataset* [44] come with free ground truth. However, enormous efforts are necessary to *realistically* model the scene complexity and effects found in reality.

3 The Variational Approach to Optical Flow Estimation

In contrast to assignment methods, variational approaches to estimating the optical flow employ a *continuous* and *dense* representation of the variables $u : \Omega \mapsto \mathbb{R}^2$. The model describing the agreement of u with the image data defines the *data term* $E_{\mathrm{D}}(u)$. It is complemented by a *regularization term* $E_{\mathrm{R}}(u)$ encoding prior knowledge about the *spatial smoothness* of the flow. Together these terms define the *energy function* $E(u)$, and estimating the optical flow amounts to finding a global minimum u, possibly constrained by a set U of admissible flow fields, and using an appropriate numerical method:

$$\inf_{u \in U} E(u)\,, \quad E(u) := E_{\mathrm{D}}(u) + E_{\mathrm{R}}(u) \tag{56}$$

$E(u)$ is non-convex in general, and hence only suboptimal solutions can be determined in practice.

Based on the variational approach published in 1981 by Horn and Schunck [29], a vast number of refinements and extensions were proposed in literature. Recent comprehensive empirical evaluations [42, 43] reveal that algorithms of this family yield best performance. Section "The Approach of Horn and Schunck" introduces the approach of Horn and Schunck as reference for the following discussion, after deriving the required linearized invariance assumption in section "Differential Constraint Equations, Aperture Problem."

Data and regularization terms designed to cope with various difficulties in real applications are presented in sections "Data Terms" and "Regularization," respectively. Section "Algorithms" gives a short overview over numerical algorithms for solving problem (56). Section "Further Extensions" addresses some important extensions of the discussed framework.

Differential Constraint Equations, Aperture Problem

All variational optical flow approaches impose an invariance assumption on some feature vector $g(x,t) \in \mathbb{R}^p$, derived from an image sequence $f(x,t)$ as discussed in section "Invariance, Correspondence Problem." Under perfect conditions, any point moving along the trajectory $x(t)$ over time t with speed $u(x,t) := \frac{\mathrm{d}}{\mathrm{d}t}x(t)$ does not change its appearance, i.e.,

$$\frac{\mathrm{d}}{\mathrm{d}t} g(x(t),t) = 0\,. \tag{57}$$

Without loss of generality, motion at $t = 0$ is considered only in what follows. Applying the chain rule and dropping the argument $t = 0$ for clarity leads to the *linearized* invariance constraint:

$$\mathrm{J}_g(x)u(x) + \partial_t g(x) = 0\,. \tag{58}$$

Validity of this approximation is limited to displacements of about 1 pixel for real data as elaborated in section "Common Aspects and Differences," which seriously limits its applicability. However, section "Multiscale" describes an approach to alleviating this restriction, and thus for now it is safe to assume that the assumption is fulfilled.

A least squares solution to (58) is given by $(S(x))^{-1}(\mathrm{J}_g^\top(x)(\partial_t g(x)))$ where

$$S(x) := \mathrm{J}_g^\top(x)\mathrm{J}_g(x). \tag{59}$$

However, in order to understand the actual information content of equation system (58), the locally varying properties of the Jacobian matrix $\mathrm{J}_g(x)$ have to be examined:

$\mathrm{rank}(\mathrm{J}_g) = 0$:	*void* constraints on $u(x)$ (for $g(x,0) = \text{const.}$);
$\mathrm{rank}(\mathrm{J}_g) = 1$:	*ill-conditioned* constraints, a single component of $u(x)$ is determined only;
$p = \mathrm{rank}(\mathrm{J}_g) = 2$:	*unique solution* $u(x) = -\mathrm{J}_g^{-1}(x)(\partial_t g(x))$;
$p > \mathrm{rank}(\mathrm{J}_g) = 2$:	*over-determined* and possibly conflicting constraints on $u(x)$, cf. Fig. 8.

In the case of gray-valued features, $g(x) = f(x) \in \mathbb{R}$, (58) is referred to as the *linearized brightness constancy constraint* and imposes *only one* scalar constraint on $u(x) \in \mathbb{R}^2$, in the direction of the image gradient $\mathrm{J}_g(x) = (\nabla g(x))^\top \neq 0$, i.e.,

$$\left\langle \frac{\nabla g(x)}{\|\nabla g(x)\|}, u(x) \right\rangle = -\frac{\partial_t g(x)}{\|\nabla g(x)\|}\,. \tag{60}$$

This limitation which only allows to determine the *normal flow* component is referred to as the *aperture problem* in the literature.

Furthermore, for real data, invariance assumptions do not hold exactly, and compliance is measured by the data term as discussed in section "Data Terms." Section "Regularization" addresses regularization terms which further incorporate regularity priors on the flow so as to correct for data inaccuracies and local ambiguities not resolved by (58).

The Approach of Horn and Schunck

The approach by Horn and Schunck [29] is described in the following due to its importance in the literature, its simple formulation, and the availability of well-understood numerical methods for efficiently computing a solution.

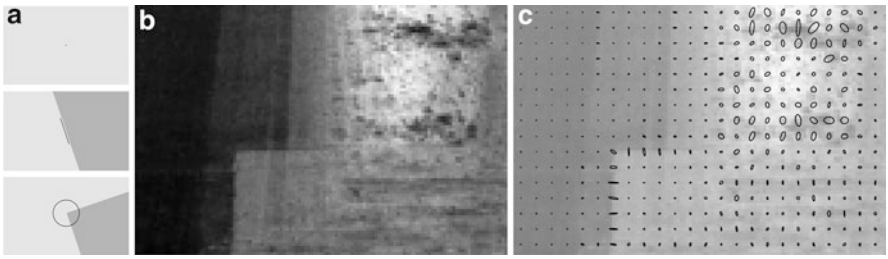


Fig. 8 Ellipse representation of $S = \mathrm{J}_g^\top \mathrm{J}_g$ as in (57) for a patch feature vector with $p \gg 2$ (see section "Patch Features"). **(a)** Three synthetic examples with J_g having (top to bottom) rank 0, 1, and 2, respectively. **(b)** Real image data with homogeneous (left) and textured (right) region, image edges, and corner (middle). **(c)** Locally varying information content (see section "Differential Constraint Equations, Aperture Problem") of the path features extracted from (**b**)

Model

Here the original approach [29], expressed using the variational formulation (56), is slightly generalized from gray-valued features $g(x) = f(x) \in \mathbb{R}$ to arbitrary feature vectors $g(x) \in \mathbb{R}^p$. Deviations from the constancy assumption in (58) are measured using a quadratic function $\rho_{\mathrm{D}} = \rho_2^2$, leading to

$$E_{\mathrm{D}}(u) = \frac{1}{2} \int_\Omega \rho_{\mathrm{D}} \left(\left\| \mathrm{J}_g(x) u(x) + \partial_t g(x) \right\|_F \right) \mathrm{d}x \,. \tag{61}$$

As for regularization, the quadratic length of the flow gradients is penalized using $\rho_{\mathrm{R}} = \rho_2^2$, to enforce smoothness of the vector field and to overcome ambiguities of the data term (e.g., aperture problem; see section "Differential Constraint Equations, Aperture Problem"):

$$E_{\mathrm{R}}(u) = \frac{1}{2\sigma^2} \int_\Omega \rho_{\mathrm{R}}(\|\mathrm{J}_u(x)\|_F) \mathrm{d}x \,. \tag{62}$$

The only parameter $\sigma > 0$ weights the influence of regularization against the data term.

Discretization

Finding a minimum of $E(u) = E_{\mathrm{D}}(u) + E_{\mathrm{R}}(u)$ using numerical methods requires discretization of variables and data in time and space. To this end, let $\{x^i\}_{i \in [n]}$ define a regular two-dimensional grid in Ω, and let $g^1(x^i)$ and $g^2(x^i)$ be the discretized versions of $g(x, 0)$ and $g(x, 1)$ of the input image sequence, respectively. Motion variables $u(x^i)$ are defined on the same grid and stacked into a vector u:

$$u(x^i) = \begin{pmatrix} u_1(x^i) \\ u_2(x^i) \end{pmatrix}, \qquad u = \begin{pmatrix} (u_1(x^i))_{i \in [n]} \\ (u_2(x^i))_{i \in [n]} \end{pmatrix} \in \mathbb{R}^{2n}. \tag{63}$$

The appropriate filter for the discretization of the spatial image gradients $\partial_l g$ strongly depends on the signal and noise properties as discussed in section "Differential Motion Estimation: Case Study (1D)." A recent comparison [11] reports that a 5-point derivative filter ($\frac{1}{12}\{-1, 8, 0, -8, 1\}$) applied to $\frac{1}{2}(g^1 + g^2)$ performs best. Temporal gradients are approximated as $\partial_t g(x^i) \approx g^2(x^i) - g^1(x^i)$.

As a result, the discretized objective function can be rewritten as

$$E(u) = \frac{1}{2}\|Du + c\|^2 + \frac{1}{2\sigma^2}\|Lu\|^2 , \tag{64}$$

using the linear operators

$$D := \begin{pmatrix} D_{1,1} & D_{1,2} \\ \vdots & \vdots \\ D_{p,1} & D_{p,2} \end{pmatrix}, \quad c := \begin{pmatrix} c_1 \\ \vdots \\ c_p \end{pmatrix}, \quad L := \begin{pmatrix} L_{1,1} \\ L_{1,2} \\ L_{2,1} \\ L_{2,2} \end{pmatrix} , \tag{65}$$

with *data* derivatives $c_j := (\partial_t g_j(x^i))_{i\in[n]}$ and $D_{j,k} := \operatorname{diag}\left((\partial_k g_j(x^i))_{i\in[n]}\right)$. The matrix operator $L_{l,k}$ applied to *variable* u approximates the spatial derivative ∂_k of the flow component u_l using the 2-tap linear filter $\{-1, +1\}$ and Neumann boundary conditions.

Solving

Objective function (64) is strictly convex in u under mild conditions [33], and thus a *global* minimum of this problem can be determined by finding a solution to $\nabla_u E(u) = 0$. This condition explicitly reads

$$(D^\top D + \sigma^{-2} L^\top L)u = -D^\top c \tag{66}$$

which is a linear equation system of size $2n$ in $u \in \mathbb{R}^{2n}$ with a positive definite and sparse matrix. A number of well-understood iterative methods exist to efficiently solve this class of problems even for large n [45].

Examples

Figure 9 illustrates the method by Horn and Schunck for a small synthetic example. The choice of parameter σ is a trade-off between smoothing out motion boundaries (see Fig. 9b) in the true flow field (Fig. 9a) and sensitivity to noise (Fig. 9d).

Probabilistic Interpretation

Considering $E(u)$ as a the log-likelihood function of a probability density function gives rise to the maximum a posteriori interpretation of the optimization problem (56), i.e.,

$$\sup_{u\in U} \mathrm{p}(u\,|g, \sigma\,) , \quad \mathrm{p}(u\,|g, \sigma\,) \propto \exp(-E(u)) . \tag{67}$$

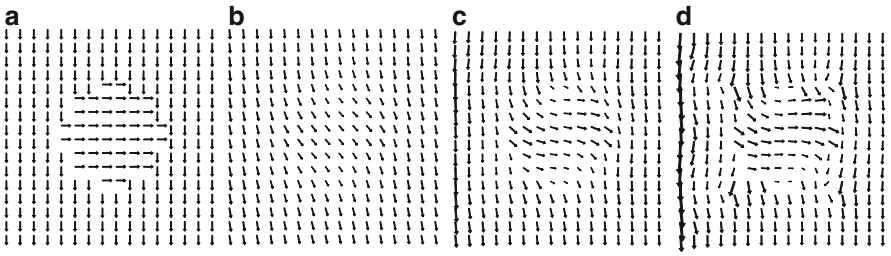


Fig. 9 **(a)** Synthetic flow field used to deform an image. **(b)–(d)** Flow field estimated by the approach by Horn and Schunck with decreasing strength of the smoothness prior

As $E(u)$ is quadratic and positive definite due to the assumptions made in section "Solving," this posterior is a Gaussian multivariate distribution

$$\mathrm{p}(u\,|g,\sigma\,) = \mathcal{N}(u;\mu,\Sigma) \tag{68}$$

with precision (inverse covariance) matrix $\Sigma^{-1} = D^\top D + \sigma^{-2} L^\top L$ and mean vector $\mu = -\Sigma^{-1} D^\top c$ that solves (66).

Examining the conditional distribution of $u^i \in \mathbb{R}^2$ allows to quantify the sensitivity of u. To this end a permutation matrix $Q = \begin{pmatrix} Q_i \\ Q_{\bar{i}} \end{pmatrix} \in \mathbb{R}^{2n\times 2n}$, $Q^\top Q = I$, is defined such that $u^i = Q_i u$. Then, fixing $Q_{\bar{i}} u = Q_{\bar{i}} \mu$ leads to

$$\mathrm{p}\left(u^i \,\middle|\, Q_{\bar{i}} u\right) = \mathcal{N}\left(\hat{\mu}^i, \hat{\Sigma}^i\right) \tag{69}$$

with $\hat{\mu}^i = Q_i \mu$ and

$$\hat{\Sigma}^i = Q_i \Sigma Q_i^\top - \left(Q_{\bar{i}} \Sigma Q_i^\top\right)\left(Q_{\bar{i}} \Sigma Q_{\bar{i}}^\top\right)^{-1}\left(Q_i \Sigma Q_{\bar{i}}^\top\right). \tag{70}$$

Using the matrix inversion theorem to invert Σ block wise according to Q and restricting the result to u_i reveals

$$Q_i \Sigma^{-1} Q_i = \left(Q_i \Sigma Q_i^\top - \left(Q_{\bar{i}} \Sigma Q_i^\top\right)\left(Q_{\bar{i}} \Sigma Q_{\bar{i}}^\top\right)^{-1}\left(Q_i \Sigma Q_{\bar{i}}^\top\right)\right)^{-1}. \tag{71}$$

Comparison of (70) to (71) and further analysis yield (for non-boundary pixels)

$$\hat{\Sigma}^i = \left(Q_i \Sigma^{-1} Q_i\right)^{-1} = \left(S^i + 4\sigma^{-2} I\right)^{-1} \tag{72}$$

with $S^i = S(x^i)$ as defined by (59). Consequently, smaller values of σ reduce the sensitivity of u^i, but some choice $\sigma > 0$ is inevitable for singular S^i.

Data Terms

Handling Violation of the Constancy Assumption

The data term as proposed by Horn and Schunck was refined and extended in literature in several ways with the aim to cope with the challenging properties of image data of real applications; see section "Basic Difficulties of Motion Estimation."

Changes of the camera viewpoint as well as moving or transforming objects may cause previously visible image features to disappear due to occlusion, or vice versa to emerge (disocclusion), leading to discontinuous changes of the *observed* image features $g(x(t), t)$ over time and thus to a violation of the invariance constraint (57).

Surface reflection properties like specular reflections that vary as the viewpoint changes, and varying emission or illumination (including shadows), also cause appearance to change, in particular in natural and outdoor scenes.

With some exceptions, most approaches do not explicitly model these cases and instead replace the quadratic distance function ρ_2^2 by the convex ℓ^2-distance or its differentiable approximation $\rho_{2,\epsilon}$, to reduce the impact of outliers in regions with strong deviation from the invariance assumption. A number of *non-convex* alternatives have been proposed in the literature, including the truncated square distance $\rho_{2,\lambda}$, which further extend this concept and are often referred to as "robust" approaches.

Another common method is to replace the constancy assumption on the image brightness by one of the more complex feature mappings $g(x, t)$ introduced in section "Invariance, Correspondence Problem," or combinations of them. The aim is to gain more descriptive features that overcome the ambiguities described in section "Differential Constraint Equations, Aperture Problem," e.g., by including color or image structure information from a local neighborhood. Furthermore, robustness of the data term can be increased by choosing features invariant to specific image transformations. For example, $g(x) = \nabla f(x)$ is immune to additive illumination changes.

Patch Features

Contrary to the strongly localized brightness feature $g(x) = f(x)$, local image patches sampled from a neighborhood $\mathcal{N}(x)$ of x,

$$g(x^i, t) = \big(f(x^j, t)\big)_{x^j \in \mathcal{N}(x^i)} \in \mathbb{R}^p, \quad p = |\mathcal{N}(x^i)| \tag{73}$$

provide much more reliable information on u in textured image regions. In fact, *local* approaches set $E_{\mathrm{R}}(u) = 0$ and rely only the information contained in the data term.

The most prominent instance introduced by Lucas and Kanade [19] chooses a Gaussian-weighted quadratic distance function,

$$\rho^2_{w^i}(z) := \left\| \mathrm{diag}(w^i)^{\frac{1}{2}} z \right\|^2, \quad w^i := (w(x^i - x^j))_{x^j \in \mathcal{N}(x^i)} \tag{74}$$

and $w(x) := \exp\left(-\|x\|^2/(2\sigma^2)\right)$. Solving the variational problem (56) decomposes into n linear systems of dimension 2 each. Furthermore, the sensitivity in terms of (72) reduces to $\hat{\Sigma}^i = (S^i)^{-1}$ and

$$S^i = \sum_{x^j \in \mathcal{N}(x^i)} w(x^i - x^j) \left(\mathrm{J}_g^\top(x^j)\mathrm{J}_g(x^j)\right) \tag{75}$$

equals the so-called *structure tensor*. At locations with numerically ill-conditioned J_g, cf. Fig. 8 and the discussion in section "Differential Constraint Equations, Aperture Problem," no flow can be determined reliably which leads to possibly sparse results. The works [46, 47] overcome this drawback by complementing this data term by a regularization term.

Multiscale

As discussed in section "Differential Motion Estimation: Case Study (1D)," the range of displacements $u(x)$ that can be accurately estimated is limited to about 1 pixel which does not conform to the larger magnitude of motion fields typically encountered in practical applications. Multiscale methods allow to remove this restriction to some extent. They implement a *coarse-to-fine* strategy for approximately determining large displacements on spatially band-limited image data and complementing flow details on finer scales.

The underlying idea is introduced by means of a multiscale representation $\{g^{[l]}\}_{l\in[n_l]}$ of image data, where $l = 0$ and $l = n_l - 1$ refer to the finest and coarsest scale, respectively. More precisely, $g^{[l]}$ is a spatially band-limited version of g with $\omega_{x,\max} < s_l\pi$ with $1 = s_0 > s_1 \ldots s_{n_l-1} > 0$. The computation is described by the following recursive scheme with $u^{[n_l]}(x) = 0$:

- $g^{[l]}(x,t) := h_l * g(x + t \cdot u^{[l+1]}, t)$
- $\delta u^{[l]} := \arg\min_u E(u)$ on data $g^{[l]}(x,t)$
- $u^{[l]}(x) := u^{[l+1]}(x) + \delta u^{[l]}(x)$

with a suitable approximation of the ideal low-pass filter h_l with frequency response

$$\hat{h}_l(\omega_x, \omega_t) \approx \begin{cases} 1 & \|\omega_x\|_\infty < s_l\pi \\ 0 & \text{otherwise} \end{cases} . \tag{76}$$

Figure 10 demonstrates the method for two simple examples.

Actual implementations further make use of the band-limited spectrum of the filtered data and subsample the data according to the Nyquist-Shannon sampling theorem, leading to a data representation referred to as *resolution pyramid*. The recursive structure allows in turn to approximate h_l by chaining filters with small support for computational efficiency.

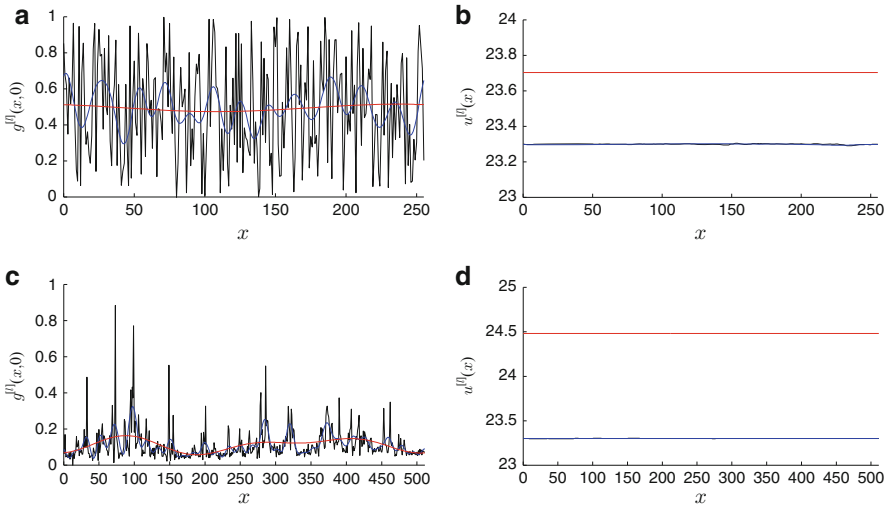


Fig. 10 Multiscale flow estimation: **(a)** an image (white noise) $g^{[l]}(x,0)$ represented at multiscale levels $l = 0$ (*black*), $l = 3$ (*blue*), and $l = 6$ (*red*) with $s_l = 2^{-l}$, i.e., band limited to $s_l\pi$. **(b)** Estimate $u^{[l]}$ (same color encoding as in **(a)**) of correct constant flow $u(x) = 23.3$ on multiscale level l. **(c)–(d)** Same as **(a)**–**(b)** for a single line of a real image (©LaVision GmbH) as found in particle image velocimetry, an optical fluid flow estimation method

Regularization

Ill-posed data terms, sensor noise, and other distortions lead to *sparse* and *locally inaccurate* flow estimates. Variational approaches allow to incorporate priors on the motion regularity by means of additional terms $E_{\mathrm{R}}(u)$. For suitable models $E_{\mathrm{D}}(u)$ and $E_{\mathrm{R}}(u)$, accuracy profits from this concept as the global solution to minimization problem (56) represents the best flow field according to both observations and priors. Furthermore, in contrast to local methods, *missing* flow information is approximately inferred according to the smoothness prior. This is in particular essential in connection with ill-posed data terms (cf. section "Differential Constraint Equations, Aperture Problem").

Regularity Priors

A number of a priori constraints $u \in U$ for flow estimation have been proposed in the literature, based on prior knowledge specific to the application domain. Examples include

- Inherent geometrical constraints induced by multi-camera setups (section "Geometrical Prior Knowledge"),
- Physical properties of flows in experimental fluid mechanics (section "Physical Prior Knowledge").

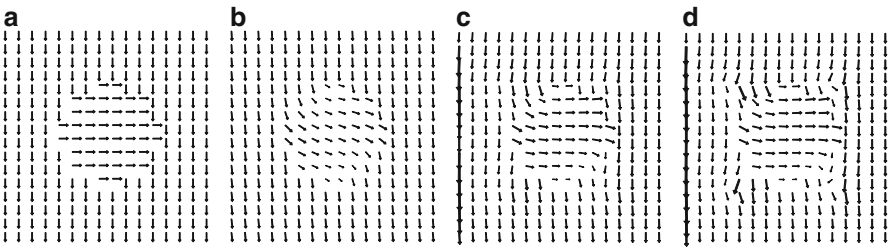


Fig. 11 **(a)** Synthetic flow field used to deform an image. **(b)–(d)** Flow field estimated by the approach by Horn and Schunck, however with $\ell_1 - TV$-regularization, with decreasing strength of the smoothness prior

Formally, strict compliance with a constraint $u \in U$ can be incorporated into the variational formulation (56) by means of the corresponding indicator function

$$E_{\mathrm{R}}(u) = \delta_U(u)\,. \tag{77}$$

In many applications, however, the set U cannot be specified precisely. Then a common approach is to replace δ_U by a smoother function measuring the distance of u to U in some sense,

$$E_{\mathrm{R}}(u) = \rho(u - \Pi_U u). \tag{78}$$

For example, the regularization term of the Horn and Schunck approach presented in section "The Approach of Horn and Schunck" may be written as

$$E_{\mathrm{R}}(u) = \|Lu\|^2 = \|u - \Pi_{\ker(L)}(u)\|_L^2 \tag{79}$$

with semi-norm $\|x\|_L := \|Lx\|$ and set $U = \ker(L)$. Generalizations of the approach of Horn and Schunck are based on the same L and modify the distance function (section "Distance Functions") or refine them to become locally adaptive and anisotropic (section "Adaptive, Anisotropic, and Nonlocal Regularization").

Further extensions replace the gradient operator in (62) and its discretization L by other operators having a larger space $U = \ker(L)$. For example, operators involving second-order derivatives $\nabla\mathrm{div}$ and $\nabla\mathrm{curl}$ have been used for flow estimation in experimental fluid dynamics [48–50] (cf. section "Physical Prior Knowledge").

Distance Functions

Occlusion of objects does not only lead to sudden changes of the projected appearance (cf. section "Data Terms") but also to *motion discontinuities* whose preservation during flow estimation is crucial in many applications and for the interpretation of videos. The penalization of large motion gradients J_u can be

reduced by replacing the quadratic distance function ρ_2^2 in (62) by convex or non-convex alternatives; see (3) for some examples.

Figure 11 demonstrates the effect of replacing the quadratic distance measure of the approach by Horn and Schunck (section "The Approach of Horn and Schunck") by $\rho_R = \rho_2$. It becomes apparent that motion discontinuities can be better resolved than with $\rho_R = \rho_2^2$ (see Fig. 9).

Adaptive, Anisotropic, and Nonlocal Regularization

A further option is to include a priori information on the *location* and *alignment* of motion discontinuities by using a spatially varying, *adaptive*, and possibly *anisotropic* norm in (62)

$$E_R(u) = \sigma^{-2} \int_\Omega \rho_R(\|J_u(x)\|_{W(x)}) \mathrm{d}x, \tag{80}$$

with $\|A\|_W := \|AW\|_F$ and (omitting the dependency on x)

$$W = \begin{pmatrix} w_1 e^1 & w_2 e^2 \end{pmatrix}. \tag{81}$$

The normalized orthogonal directions $e^1, e^2 \in \mathbb{R}^2$ point *across* and *along* the assumed motion boundary, respectively. The positive eigenvalues w_1 and w_2 control relative penalization of flow changes in the according direction.

A common assumption made in literature, e.g., [36, 51], is that image edges and flow discontinuities coincide and facilitate changes of $u(x)$ *across* the assumed boundary e_1. For general features $g(x)$, the notion of image edge is here defined by choosing e^1 and e^2 as the normalized direction e of the largest and smallest change of $\|J_g e\|$, respectively, given by the eigenvectors of $S = J_g^\top J_g$. The associated eigenvalues $\lambda_1 \geq \lambda_2 \geq 0$ of S control the strength of smoothness by setting $w_i = 1 - \rho(\lambda_i)$, $i = 1, 2$ and suitable increasing $\rho(x) \in [0, 1]$ with $\rho(0) = 0$. This defines an *anisotropic* and *image-driven* regularization. Note that for the gray-valued case $g(x) = f(x) \in \mathbb{R}$, the formulation simplifies to $e^1 = \|\nabla g\|^{-1} \nabla g$, $\lambda_1 = \|\nabla g\|^2$, and $\lambda_2 = 0$. The class of *flow-driven* approaches replace the dependency on $g(x)$ of the terms above by the flow $u(x)$ to be estimated. This nonlinear dependency can be taken into account without compromising convexity of the overall variational approach [36].

While the approaches so far measure *locally* the regularity of flows u, approaches such as [52] adopt nonlocal functionals for regularization developed in other contexts [53–55] for optical flow estimation. Regularization is then more generally based on the similarity between *all* pairs $(u(x), u(x'))$ with $x, x' \in \Omega$, weighted by mutual position and feature distances.

Further Extensions

Three extensions of the basic variational approach are sketched: a natural extension of spatial regularizers to the spatiotemporal domain (section "Spatiotemporal Approach"), regularization based on the two-view geometry (cf. section "Two-View Geometry, Assignment and Motion Fields") and relative rigid motions for computer vision applications (section "Geometrical Prior Knowledge"), and a case study of PDE-constrained variational optical flow estimation in connection with imaging problems in experimental fluid dynamics (section "Physical Prior Knowledge").

Spatiotemporal Approach

The preceding discussion reduced the motion estimation problem to determining displacements between *two* image frames only and thus ignored consistencies of the flow *over time*. Although in many applications recording rates are fast compared to dynamical changes due to modern sensors, only few approaches exploit this fact by introducing *temporal smoothness priors*.

The work [35] proposed to process a batch of image frames *simultaneously* and to extend the flow field domain along the time axis $u : \Omega \times [0, T] \mapsto \mathbb{R}^2$. While data terms are independently imposed for each time t, the smoothness prior is extended by a temporal component to

$$E_{\mathrm{R}}(u) := \int_{\Omega \times [0,T]} \rho_{\mathrm{R}}(\|\mathrm{J}_{u,t}(x,t)\|_W) \mathrm{d}x \mathrm{d}t \,. \tag{82}$$

Here, $\mathrm{J}_{u,t}$ represents the spatiotemporal derivatives, and $\rho_{\mathrm{R}}\|\cdot\|_W$ is a three-dimensional extension of the *anisotropic, flow-driven* distance function discussed in section "Adaptive, Anisotropic, and Nonlocal Regularization." It allows to account for *small* position changes of moving objects between consecutive frames within the support of the regularization term ($\leq$1 px) by supporting smoothness *along* an assumed trajectory.

Larger displacements, however, require matching of temporally associated regions, e.g., using a multiscale framework (section "Multiscale"), but then enable to regularize smoothness of *trajectories* over multiple frames as proposed in [56].

Online methods are an appealing alternative whenever processing a batch of image frames is not feasible due to resource limitations. This approach is addressed in section "Probabilistic Modeling and Online Estimation."

Geometrical Prior Knowledge

In applications with a perspective camera as image sensor, the geometrical scene structure strongly determines the observed optical flow (section "Two-View Geometry, Assignment and Motion Fields"). This section briefly addresses the most common assumptions made and the constraints that follow.

Often, a *static scene* assumption is made, meaning that all visible scene points have zero velocity with respect to a world coordinate system. Then the observed motion is only induced by the camera moving in the scene. Using the notation

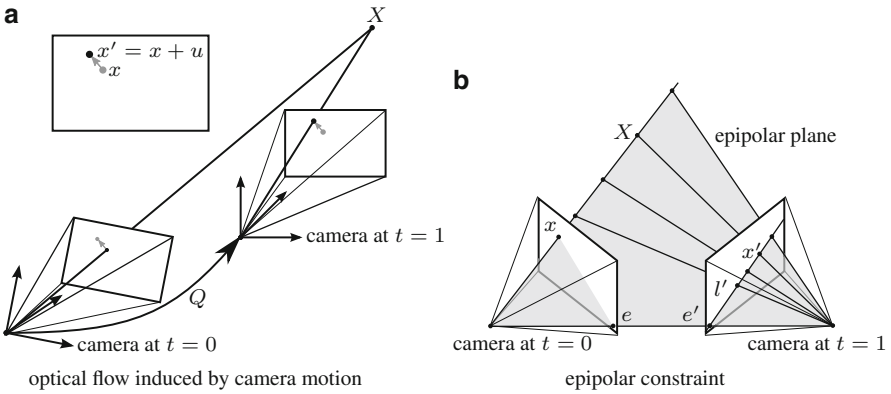


Fig. 12 **(a)** Relative motion $Q \in \mathsf{SE}(3)$ of the camera w.r.t. a world coordinate system causes the projection x of a static scene point $X = z(x)y$ to move from x to $x' = x + u(x)$ in the image plane. **(b)** Any two projections x, x' of a scene point X are related by the essential matrix E as in (84), defining an *epipolar plane* and their projection, the *epipolar lines* defined by $\{x, e\}$ and $\{x', e'\}$ in the image plane at $t = 0$ and $t = 1$, respectively

introduced in section "Two-View Geometry, Assignment and Motion Fields," the camera motion is denoted by $Q \in \mathsf{SE}(3)$ (cf. Fig. 12a), parametrized by rotation $R \in \mathsf{SO}(3)$ and translation $h \in \mathbb{R}^3$, so that any scene point $Y \in$ is transported to $Y' \simeq Q^{-1}Y$.

The following discussion of common setups and their implications on the observed motion implicitly assumes that the scene point is visible in both frames. Using assumption (41) for the internal camera parameters allows to work with normalized coordinates (32). The point corresponding to x is denoted by x', due to (42).

Static scene, general motion Let the depth map $z(x) : \Omega \mapsto \mathbb{R}$ parametrize the scene point $X := z(x) \binom{x}{1}$ visible at x in the camera plane in the first frame. Then the projected correspondences are given in homogeneous coordinates by

$$y' \simeq PQ^{-1}Y = R^\top \left(z(x) \binom{x}{1} - h\right) , \tag{83}$$

see Fig. 12a for an illustration. Figure 13 shows the optical flow field $u(x)$ conforming to constraint (83) for a real application.

It is possible to eliminate the dependency on $z(x)$ that typically is unknown by means of the *essential matrix* $E := R^\top [h]_\times$, leading to the *epipolar constraint*

$$(y')^\top E y = 0 , \tag{84}$$

as illustrated by Fig. 12b. This gives rise to an orthogonal decomposition [58] of an observed correspondence $\hat{x}'$ into

Fig. 13 **(a)** A single frame from an image sequence recorded by a camera moving forward through an approximately static scene. **(b)** Optical flow estimated using the parametrization $u(x) = u(x; Q, z(x))$ according to (83) and global optimization for $Q \in \mathsf{SE}(3)$, $z \in \mathbb{R}^n$; see [57] for details. Displacement length and direction are encoded by saturation and hue, respectively; see color code on the right. **(c)** Estimated depth parameter $z(x)$ using the color code on the right. Scene structure is more evident in this representation, and therefore the spatial smoothness prior on the flow was formulated as regularization term on the *depth* $z(x)$ instead of displacements $u(x)$

$$\hat{x}' = \hat{x}'_e + \hat{x}'_\perp \tag{85}$$

with $\hat{x}'_e$ fulfilling (84) and orthogonal deviations $\hat{x}'_\perp$.

Even without knowing a priori (R, h), Eq. (84) provides a valuable prior: Valgaerts et al. [59] propose joint computation of the fundamental matrix F related to E by (40) and optical flow constrained via (84). They show that estimation of F is more stable and that flow accuracy is significantly increased.

Static scene, coplanar camera motion If the camera translates parallel to the image plane only, i.e., $R = I$ and $h = \binom{b}{0}$ with $b \in \mathbb{R}^2$, the observed flow is constrained to a locally varying *one-dimensional* subspace parametrized by the inverse depth:

$$u(x) = z^{-1}(x) b \,. \tag{86}$$

Stereoscopic camera setups fulfill the static scene assumption as they can be interpreted as an instantaneous camera motion with baseline $\|b\|$. For details see, e.g., [5].

Planar and static scene, general camera motion In applications where the scene can be (locally) approximated by a plane such that $\langle n, X \rangle - d = 0$ for all space points X with plane parameters d, n as in (46), all correspondences fulfill

$$y' \simeq Hy, \; H = R^\top \left(I - \frac{1}{d} h n^\top \right) , \tag{87}$$

where $H \in \mathbb{R}^{3\times 3}$ defines a *homography* – cf. Eq. (49).

Physical Prior Knowledge

Imaging of dynamic phenomena in natural sciences encounters often scenarios where physical prior knowledge applies. Examples include particle image velocime-

try [60] or Schlieren velocimetry [61], where the motion of fluids is observed, that is, governed by physical laws. While local methods such as cross-correlation methods are commonly used to evaluate the obtained image sequences [60, 62], variational approaches [63–65] provide a more appropriate mathematical framework for exploiting such prior knowledge and the estimation of physically consistent optical flows.

For instance, the Helmholtz decomposition of vector fields enables to define regularizers in terms of higher-order partial flow derivatives in a natural way [48–50]. Constraints like incompressibility can be enforced as hard or soft constraints using advanced methods of convex programming, to cope with imaging imperfections. Conversely, flow field estimates obtained by other image processing methods can be denoised so as to restore physically relevant structure [66].

A particularly appealing approach exploits directly some equation from fluid dynamics that governs the flow as state of the physical system which is observed through an imaging sensor [67, 68]. The state is regarded as hidden and only observable through the data of an image sequence that depicts the velocity of some tracer suspended in the fluid. The variational approach of fitting the time varying state to given image sequence data results in a PDE-constrained optimization or distributed parameter control problem, respectively.

As example the approach [67] is sketched based on the Stokes system

$$-\mu \Delta u + \nabla p = f_\Omega \qquad \text{in } \Omega, \tag{88a}$$

$$\operatorname{div} u = 0 \qquad \text{in } \Omega, \tag{88b}$$

$$u = f_{\partial\Omega} \qquad \text{on } \partial\Omega, \tag{88c}$$

that for given f_Ω, $f_{\partial\Omega}$ with $\int_{\partial\Omega} \langle n, f_{\partial\Omega}\rangle \mathrm{d}s = 0$ (n denotes the outer unit normal of the Lipschitz domain Ω) has a unique solution u, p under classical assumptions [69, Ch. I]. Here f_Ω, $f_{\partial\Omega}$ are not regarded as given data but as control variables, to be determined so that the flow u not only satisfies (88) but fits also given image sequence data. To achieve the latter, both the state variables u, p and the control variables f_Ω, $f_{\partial\Omega}$ are determined by minimizing in the two-dimensional case $d = 2$ the objective

$$E(u, p, f_\Omega, f_{\partial\Omega}) = E_{\mathrm{D}}(u) + \alpha \int_\Omega \rho_2^2(f_\Omega)\mathrm{d}x + \gamma \int_{\partial\Omega} \rho_2^2\big(\langle n^\perp, \nabla f_{\partial\Omega}\rangle\big)\mathrm{d}s, \quad \alpha, \gamma > 0. \tag{89}$$

The first term $E_{\mathrm{D}}(u)$ denotes a data term of the form (61), and the remaining two terms regularize the control variables so as to make the problem well posed.

For related mathematical issues (e.g., constraint qualification and existence of Lagrange multipliers), see [70, Ch. 6] and [71, Ch. 1] and furthermore [70, 72] for related work outside the field of mathematical imaging based on the general Navier-Stokes system.

Algorithms

The choice of an optimization method for numerically minimizing the functional (56) depends on the specific formulation of the terms E_D and E_R involved. Suitable methods can be broadly classified into

- Algorithms for minimizing *smooth convex* functionals,
- Algorithms for minimizing *non-smooth convex* functionals,
- Algorithms for locally minimizing *non-convex* functionals.

In view of the typical multiscale implementation of the data term (section "Multiscale") that enables a quadratic approximation at each resolution level, this classification is applied to the regularizer E_R only, and each class is discussed in turn in the sections to follow. The reader should note that convex non-quadratic data terms, as discussed in section "Handling Violation of the Constancy Assumption," can be handled in a similar way as the convex non-smooth regularizer below, and a number of closely related alternatives exist (e.g., [73]). Since convex programming has been extensively studied in the literature, the following presentation is confined to representative case studies that illustrate in each case the underlying idea and application of a general principle.

Smooth Convex Functionals

It is useful to distinguish quadratic and non-quadratic functionals. The approach of Horn and Schunck (section "The Approach of Horn and Schunck") is a basic representative of the former class. Solving the corresponding linear positive definite sparse system can be efficiently done by established methods [45]. More sophisticated implementations are based on numerical multigrid methods [74]. These are optimal in the sense that runtime complexity $O(n)$ linearly depends on the problem size n. Dedicated implementations run nearly at video frame rate on current PCs.

For more general data-dependent quadratic regularizers and especially so for non-quadratic convex regularizers (cf. section "Adaptive, Anisotropic, and Nonlocal Regularization" and [36]), multigrid implementation that achieves such runtimes requires some care. See [12, 13, 75] for details and [76] for a general exposition.

Non-smooth Convex Functionals

This class of optimization problems has received considerable attention in connection with mathematical imaging, inverse problems, and machine learning and in other fields during the recent years, due to the importance of non-smooth convex sparsity enforcing regularization. See [77] for a recent overview.

The *total variation* regularizer

$$\begin{aligned} E_{\mathrm{R}}(u) = \mathrm{TV}(u) &:= \sup_{v \in \mathcal{D}} - \int_{\Omega} \langle u, \operatorname{Div} v \rangle \mathrm{d}x, \\ \mathcal{D} &:= \{ v \in C_0^\infty(\Omega; \mathbb{R}^d)^d \colon \|v(x)\|_F \leq 1, \ \forall x \in \Omega \}, \\ \operatorname{Div} v &= (\operatorname{div} v^1, \ldots, \operatorname{div} v^d)^\top \end{aligned} \tag{90}$$

is a basic representative of the class of non-smooth convex functionals and appropriate to expose a general strategy of convex programming that is commonly applied: problem splitting into subproblems for which the proximal mapping can be efficiently evaluated.

The simplest *anisotropic* discretization of (90) that is particularly convenient from the viewpoint of convex programming reads

$$\sum_{ij \in E(G)} \sum_{k \in [d]} |u_k(x^i) - u_k(x^j)|, \tag{91}$$

where $\{x^i\}_{i \in [n]}$ are the locations indexed vertices $V = [n]$ of a grid graph $G = (V, E)$ in Ω, and $E = E(G)$ are the corresponding edges connecting adjacent vertices resp. locations along the coordinate axes. Defining the vector

$$z \in \mathbb{R}^{d \times |E(G)|}, \quad z_{k,ij} = u_k(x^i) - u_k(x^j) \tag{92}$$

leads to the reformulation of (91)

$$\|z\|_1, \qquad Lu = z \tag{93}$$

where the linear system collects all equations of (92). As a consequence, the overall discretized problem reads

$$\min_{u,z} E_{\mathrm{D}}(u) + \alpha \|z\|_1 \quad \text{subject to} \quad Lu - z = 0, \qquad \alpha > 0 \tag{94}$$

to which the ADMM approach [78] can be applied that entails a sequence of partial minimizations of the augmented Lagrangian corresponding to (94),

$$L_\lambda(u, z, w) = E_{\mathrm{D}}(u) + \alpha \|z\|_1 + \langle w, Lu - z \rangle + \frac{\lambda}{2} \|Lu - z\|^2. \tag{95}$$

Specifically, with some parameter value $\lambda > 0$ and multiplier vector w, the three-step iteration

$$u^{k+1} = \arg\min_u E_{\mathrm{D}}(u) + \langle w^k, Lu \rangle + \frac{\lambda}{2} \|Lu - z^k\|^2, \tag{96a}$$

$$z^{k+1} = \arg\min_z \alpha\|z\|_1 - \langle w^k, z\rangle + \frac{\lambda}{2}\|Lu^{k+1} - z\|^2, \tag{96b}$$

$$w^{k+1} = w^k + \lambda(Lu^{k+1} - z^{k+1}), \tag{96c}$$

is iteratively applied for $k = 0, 1, 2, \ldots$, with arbitrary initializations z^0, q^0, until a suitable termination criterion is met [78, Section 3.3.1].

Assuming a quadratic form or approximation of $E_D(u)$ at some resolution level (section "Multiscale"), subproblem (96a) amounts to solve a sparse positive definite linear system similar to the basic approach of Horn and Schunck, to which a multigrid solver can be applied as discussed above. Subproblem (96b) amounts to computing the proximal mapping for the ℓ^1-norm and hence to perform a simple shrinkage operation. See [79, 80] for corresponding surveys.

Non-convex Functionals

Similar to the preceding non-smooth convex case, approaches are of interest that can be conducted by solving a sequence of *simple* subproblems efficiently. Clearly, convergence to a *local* minimum can be only expected. In contrast to the simpler convex cases above, the absence of parameters is preferable that would have to be set properly, to ensure convergence to some local minimum for any initialization. For example, Lipschitz constants of gradients are rarely known in practice, and setting corresponding parameters savely enough will unduly slow down convergence even for smooth problems.

A general strategy will be outlined next and its application to the non-convex extension of the regularizer (91), using the distance function (3f),

$$\sum_{ij \in E(G)} \rho_{2,\lambda}\big(u(x^i) - u(x^j)\big). \tag{97}$$

In order to illustrate graphically the non-convexity of this regularizer from the viewpoint of optimization, consider three summands of the "fully" anisotropic version of (97),

$$\sum_{ij \in E(G)} \sum_{k \in [d]} \rho_{2,\lambda}\big(u_k(x^i) - u_k(x^j)\big). \tag{98}$$

defined on edges that meet pairwise in a common vertex,

$$\rho_{2,\lambda}\big(u_k(x^{i_1}) - u_k(x^{i_2})\big) + \rho_{2,\lambda}\big(u_k(x^{i_2}) - u_k(x^{i_3})\big) + \rho_{2,\lambda}\big(u_k(x^{i_3}) - u_k(x^{i_4})\big). \tag{99}$$

Setting for simplicity and w.l.o.g. $u_k(x^{i_1}) = u_k(x^{i_4}) = 0$ to obtain a function of two variables $u_k(x^{i_2})$, $u_k(x^{i_3})$ results in the corresponding graph depicted by Fig. 14. It illustrates the presence of many non-strict local minima and that the design of a convergent minimization algorithm is not immediate.

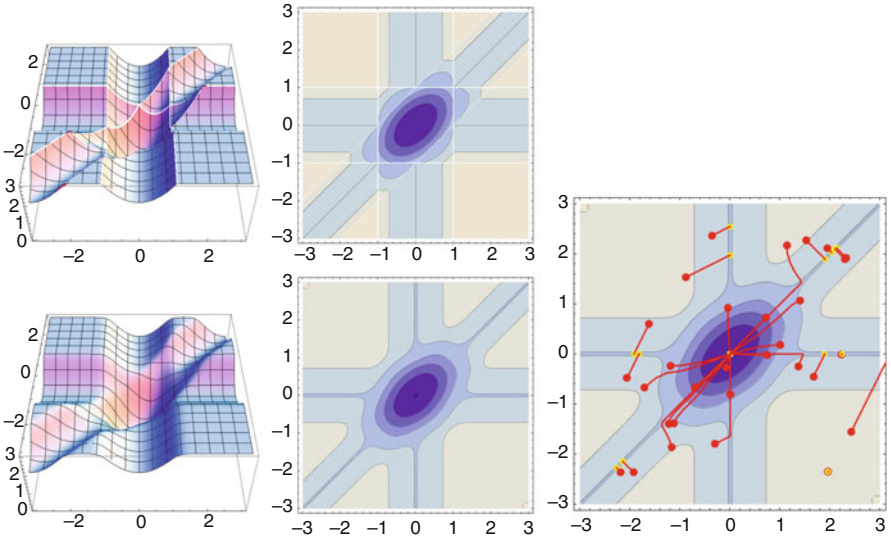


Fig. 14 *Top row, left*: two different illustrations of the non-convex, non-smooth objective (99). *Bottom row, left*: the objective (99) smoothed by replacing the distance function $\rho_{2,\lambda}$ by $\rho_{2,\lambda,\varepsilon}$ with $\varepsilon = 0.2$, as defined by (3g). *Right panel*: sequences of iterates generated by (103) for 30 random points $(z_2, z_3)^\top$ (initial and final iterates are marked with *red* and *yellow*, respectively). The regularizer enforces fitting of the components z_2, z_3 to the data $z_1 = z_4 = 0$ as well as $z_2 = z_3$. It is robust in the sense that components that are too distant to either of these criteria are not affected accordingly

Next consider a single summand $\rho_{2,\lambda}(z_i - z_j)$ of (98) with two scalar variables denoted by z_i and z_j for simplicity. This function can be decomposed into the difference of two proper, lower semicontinuous (lsc), convex functions g and h:

$$\begin{aligned}\rho_{2,\lambda}(z_i - z_j) &= \tau(z_i - z_j)^2 - \Big(\tau(z_i - z_j)^2 - \rho_{2,\lambda}(z_i - z_j)\Big) \\ &=: g(z_i, z_j) - h(z_i, z_j), \qquad \tau > 1.\end{aligned} \tag{100}$$

Applying this decomposition to each term of (98) yields

$$g(u) - h(u) \tag{101}$$

with $g(u) = \tau\|Lu\|^2$ as in (64), and with $h(u)$ equal to the sum of all edge terms of the form $h(u_k(x^i), u_k(x^j))$, $ij \in E$, $k \in [d]$, given by (100).

DC-programming (DC stands for Difference-of-Convex functions [81]) amounts to locally minimize (101) by solving a sequence of convex problems, defined by the closed affine majorization of the concave part $-h$,

$$u^{k+1} = \arg\min_u g(u) - \big(h(u^k) + \langle v^k, u - u^k\rangle\big), \qquad v^k \in \partial h(u^k), \tag{102}$$

where $\partial h(u^k)$ denotes the subdifferential of h at u^k. This two-step iteration in terms of (u^k, v^k) converges under mild conditions [82]. Smoothing the problem slightly by replacing the distance function $\rho_{2,\lambda}$ by $\rho_{2,\lambda,\varepsilon}$ defined by (3g), and replacing accordingly h by h_ε, yields $v^k = \nabla h_\varepsilon(u^k)$ and hence turns (102) into the sequence of problems

$$u^{k+1} = \arg\min_u g(u) - \langle \nabla h_\varepsilon(u^k), u - u^k \rangle. \tag{103}$$

Taking additionally into account the data term $E_D(u)$ and assuming it (or its approximation) has quadratic form at some resolution level (section "Multiscale"), solving (103) amounts to a sequence of Horn and Schunck type problems to which numerical multigrid can be applied, due to the simple form $g(u) = \tau \|Lu\|^2$. Not any single parameter, e.g., for selecting the stepsize, has to be set in order to ensure convergence, and available code for a variational method can be directly applied. The price to pay for this convenience is a moderate convergence rate.

Figure 14 illustrates the beneficial effect of smoothing and robustness of the non-convex regularizer: Only the components of points $\binom{z_2}{z_3}$ that are close enough to the data $z_1 = u_k(x^{i_1}) = z_4 = u_k(x^{i_4}) = 0$, as specified by λ, are fitted to these data. For distant points with $z_2 \approx z_3$, regularization enforces $z_2 = z_3$ or does not affect them at all if $|z_2 - z_3|$ is large.

Applying the scheme (103) to (97) instead of (98) is straightforward. This does not affect $g(u)$ but merely ∇h_ε in (103), due to replacing the scalar variables in (100) by the corresponding vectors.

4 The Assignment Approach to Optical Flow Estimation

In this section approaches to determining the assignment field $u(x,t)$ (4) are considered that establish the correspondence (2) of a given feature mapping $g(x,t)$ in two given images.

The following sections conform to a classification of these approaches. Both the scope and the application areas associated with each class of approaches overlap with the variational approach of Sect. 3 but otherwise differ. The presentation focuses on the former aspects and the essential differences, whereas an in-depth discussion of the latter aspects is beyond the scope of this survey.

Section "Local Approaches" discusses local approaches to the assignment problem, whereas the remaining three sections are devoted to global approaches. In section "Assignment by Displacement Labeling" the correspondence problem is reformulated as a labeling problem so that methods for solving the *Maximum A Posteriori (MAP)* problem with the corresponding *Markov random field (MRF)* model can be applied. Assignment by *variational image registration* is briefly considered in section "Variational Image Registration."

Local Approaches

Key feature of the class of assignment approaches is the restriction of the set of feasible assignment fields $u(x)$ to a *finite* set. This set is defined by restricting at each location $\{x^i\}_{i\in[n]} \in \Omega$ the range of $u(x^i) \in \mathcal{U}(x^i)$ to a finite set $\mathcal{U}(x^i)$.

Local approaches determine the optimal $u(x^i)$ independently, i.e., they solve for each $i \in [n]$

$$u(x^i) \in \operatorname*{arg\,min}_{u\in\mathcal{U}(x^i)} \rho\big(g(x^i,t), g(x^i+u, t+\delta t)\big). \tag{104}$$

The usually small sets $|\mathcal{U}(x^i)|$ allow *exhaustive search* to find an optimal solution. Thus, the general distance function $\rho(\cdot,\cdot)$ is not required to be convex or differentiable and allows for more involved formulations.

Since local methods do not make use of (nonlocal) spatial smoothness priors w.r.t. u, they require – and, in fact, solely rely on – discriminative features, typically derived from local images patches also used by local *variational* methods; see (73):

$$g(x^i,t) = \big(f(x^j,t)\big)_{x^j\in\mathcal{N}(x^i)} \in \mathbb{R}^p, \quad p = |\mathcal{N}(x^i)| \tag{105}$$

with some neighborhood $\mathcal{N}(x^i)$, e.g., a square region.

In the following some common choices for ρ are addressed. For brevity, the discussion omits references to x^i and some fixed $u = u(x^i)$ and puts $g^1 := g(x^i,t^1)$, $g^2 := g(x^i+u,t^2)$ with $t^2 = t^1 + \delta t$.

Template-based matching methods compare a template g^1 pixel wise to a potential match g^2 and derive some similarity measure from it. Direct comparison of gray values,

$$\rho(g^1,g^2) = \rho(g^1-g^2) \tag{106}$$

is usually avoided in favor of distance functions which are invariant to brightness or geometric changes. Two popular choices are:

- The *normalized cross-correlation* [83] derives patch features which are invariant to global additive and multiplicative changes of g by defining

$$\overline{g}^k = \frac{g^k - \mu(g^k)}{\sigma(g^k)}, \qquad k = 1,2 \tag{107}$$

  with mean $\mu(g^k)$ and standard deviation $\sigma(g^k)$ of samples $\{g^k_j\}_{j\in[p]}$. Then the distance function is defined as

$$\rho_{\mathrm{NCC}}(g^1,g^2) = 1 - \frac{1}{p}\langle \overline{g}^1, \overline{g}^2\rangle = \frac{1}{2p}\rho_2^2(\overline{g}^1 - \overline{g}^2) \tag{108}$$

where the last equation follows from $\langle \overline{g}^k, \overline{g}^k \rangle = p\sigma^2(\overline{g}^k) = p$.

- The *Census transform* creates binary descriptors

$$\overline{g}^k = \left(\psi_{\mathbb{R}_+}(g_j^k - m^k)\right)_{j\in[p]} \in \{0,1\}^p, \qquad k = 1,2 \tag{109}$$

with $m^k := g(x^i, t^k)$, which approximate directional derivatives [84] and measures the Hamming distance

$$\rho_{\mathrm{CT}}(g^1, g^2) = \rho_1\left(\overline{g}^1 - \overline{g}^2\right) = \|\overline{g}^1 - \overline{g}^2\|_1 . \tag{110}$$

This transformation is in particular invariant to any strictly monotonically increasing transformation $\gamma\colon \mathbb{R} \mapsto \mathbb{R}$ uniformly applied to all components of g^1 and g^2.

Histogram-based methods relax the pixel-by-pixel comparison in (106) to achieve additional invariance to geometric transformations.

Exemplarily, a method frequently used in medical images registration [85] and stereo disparity estimation [86] is detailed. It uses the concept of *mutual information* [87] to measure distances between gray-value probability distributions $\hat{\mathrm{p}}_k(f; g^k)$, $k = 1,2$, determined as kernel density estimates [88] from the samples $\{g_j^k\}_{j\in[p]}$. Their entropies are given by

$$\mathrm{H}(\hat{p}_k; g^k) = -\int \hat{\mathrm{p}}_k(f; g^k)\, \log \hat{\mathrm{p}}_k(f; g^k)\, \mathrm{d}f, \qquad k = 1,2. \tag{111}$$

The joint distribution $\hat{\mathrm{p}}_{1,2}(f^1, f^2; g^1, g^2)$ is defined accordingly with joint entropy

$$\mathrm{H}(\hat{\mathrm{p}}_{1,2}; g^1, g^2) = -\int \hat{\mathrm{p}}_{1,2}(f^1, f^2; g^1, g^2)\, \log \hat{\mathrm{p}}_{1,2}(f^1, f^2; g^1, g^2)\, \mathrm{d}f^1 \mathrm{d}f^2. \tag{112}$$

Then the mutual information defines the distance function

$$\rho_{\mathrm{MI}}(g^1, g^2) = \mathrm{H}(\hat{p}_1; g^1) + \mathrm{H}(\hat{p}_2; g^2) - \mathrm{H}(\hat{\mathrm{p}}_{1,2}; g^1, g^2) \tag{113}$$

which shows some robustness against rotation, scaling, and illumination changes.

Complex approaches such as *scale-invariant feature transform (SIFT)* [3] and *speeded-up robust features (SURF)* [4] combine several techniques including histogram of orientations and multiple resolution to optimize robustness, reliability, and speed.

Assignment by Displacement Labeling

Consider again sets $\mathcal{U}(x^i)$ of assignment vectors as discussed in section "Local Approaches." In contrast to local approaches presented in the previous section, this section is devoted to methods that *simultaneously* select vectors $u(x^i) \in \mathcal{U}(x^i)$ for all locations x^i, $i \in [n]$, based on optimization criteria that evaluate desired properties of assignment fields u. The feasible set of u is denoted by $\mathcal{U} := \cup_{i \in V} \mathcal{U}(x^i)$. It will be convenient to index locations $\{x^i\}_{i \in V} \in \Omega$ by vertices $i \in V = [n]$ of a graph $G = (V, E)$.

As a consequence of the twofold discretization of both the underlying domain $\Omega \subset \mathbb{R}^d$ and the range of $u(x)$, it makes sense to associate with each location x^i an integer-valued variable

$$\ell_i := \ell(x^i) \in [m_i], \qquad m_i := |\mathcal{U}(x^i)|, \tag{114}$$

whose value determines the assignment vector $u(x^i) \in \mathcal{U}(x^i)$. This separates the problem formulation in terms of the *labeling field* $\ell := \{\ell_i\}_{i \in V}$ from the set of assignment vectors $\mathcal{U}$ that may vary, as is further discussed below.

Analogous to objectives (56) of variational approaches, a functional as criterion for labelings ℓ defines an approach,

$$\begin{aligned} J(\ell; \mathcal{U}) &= J_{\mathrm{D}}(\ell) + J_{\mathrm{R}}(\ell) \\ &= \sum_{i \in V} \varphi_i(\ell_i; \mathcal{U}) + \sum_{ij \in E} \varphi_{ij}(\ell_i, \ell_j; \mathcal{U}), \end{aligned} \tag{115}$$

together with an algorithm for determining an assignment field u in terms of a minimizing labeling field ℓ. For instance, in view of a data term like (61), a reasonable definition of the function $\varphi_i(\cdot; \mathcal{U})$ of (115) is

$$\varphi_i(\ell_i; \mathcal{U}) = \rho_{\mathrm{D}}\left(\left\| \mathrm{J}_g(x^i) u_{\ell_i} + \partial_t g(x^i) \right\|_F \right), \quad u_{\ell_i} \in \mathcal{U}(x^i), \quad \ell_i \in [m_i] \tag{116}$$

where ℓ_i enumerates all possible assignment vectors u_{ℓ_i} at x^i. However, getting back to the differences to the differential approach addressed in section "Common Aspects and Differences," a major motivation of formulation (115) is to disregard partial derivatives of the feature map involved in differential variational approaches (section "Differential Constraint Equations, Aperture Problem") and hence to avoid the corresponding limitations discussed in sections "Differential Motion Estimation: Case Study (1D)" and "Multiscale." Rather, data terms J_{D} are directly defined by setting up and evaluating *locally* possible assignments $u_{\ell_i} \in \mathcal{U}(x^i)$ that establish a correspondence between local features (2), extracted from the given image pair, and by defining costs $\varphi_i(\ell_i; \mathcal{U})$ accordingly. Notice that no smoothness of φ_i is required – any distance discussed in section "Local Approaches" may be employed as in (116). For a discussion of the distance (113) in this connection, see [89].

The same remarks apply to the definition of J_{R} in (115). A common choice in the literature however is the discrete version of the non-convex regularizer (97)

$$\varphi_{ij}(\ell_i, \ell_j; \mathcal{U}) = \rho_{2,\lambda}(u_{\ell_i} - u_{\ell_j}; \mathcal{U}). \tag{117}$$

The reader should notice that the *non-convex* regularizer (97) has been replaced by the *combinatorial* version (117). Likewise, the *non-convex* data term (61) has been replaced by the *discrete-valued* term (116). More generally, the problem related to the variational approach to cope with the non-convexity of the data term (section "Multiscale") by means of a multiscale implementation (section "Multiscale"), and with the non-convexity of the overall functional by computing a "good" local minimum (section "Non-convex Functionals"), has been replaced by the combinatorial problem to determine an optimal assignment by minimizing (115). This problem is known in the literature as *maximum a posteriori (MAP) problem* w.r.t. the discrete probabilistic graphical model

$$\mathrm{p}_G(\ell; \mathcal{U}) = \frac{1}{Z} \exp\big(-J(\ell; \mathcal{U})\big), \qquad Z = \sum_{\ell} \exp\big(-J(\ell; \mathcal{U})\big), \tag{118}$$

that is the problem to compute the *mode* $\arg\max \mathrm{p}_G(\ell; \mathcal{U})$ of the *Markov random field* p_G defined on the undirected graph G. See [90, 91] for background and further details.

Many past and current research activities are devoted to this problem, across various fields of computer science and applied mathematics. Approaches range from integer programming techniques to various convex relaxations and combinations thereof. To get a glimpse of the viewpoint of polyhedral combinatorics on the problem to minimize (115), consider a single summand $\varphi_i(\ell_i; \mathcal{U})$ and define the vector

$$\theta^i \in \mathbb{R}^{m_i}, \qquad \theta^i_{\ell_i} := \varphi_i(\ell_i; \mathcal{U}), \qquad \ell_i \in [m_i], \tag{119}$$

whose components specify the finite range of the function φ_i. Then the problem of determining $\overline{\ell}_i$ corresponding to the minimal value of $\varphi_i(\ell_i; \mathcal{U})$ can be rewritten as

$$\min_{\mu^i \in \Delta_{m_i}} \langle \theta^i, \mu^i \rangle, \tag{120}$$

which is a *linear program (LP)*. Clearly, for general data defining θ^i by (119), the vector $\overline{\mu}^i$ minimizing (120) is a vertex of the simplex Δ_{m_i} corresponding to the indicator vector $\overline{\mu}^i = (0, \ldots, 0, 1, 0, \ldots, 0)^\top$ of the value $\overline{\ell}_i$. This reformulation can be applied in a straightforward way to the overall problem of minimizing (115), resulting in the LP

$$\min_{\mu \in \mathcal{M}_G} \langle \theta, \mu \rangle, \tag{121}$$

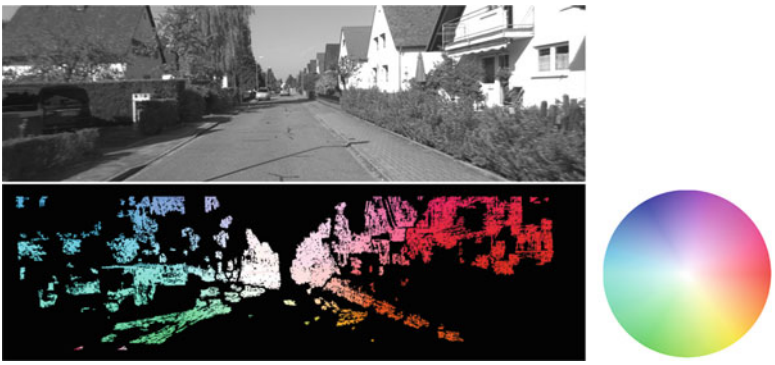


Fig. 15 *Top* Frame of a sequence, taken with a fast moving camera from the KITTI benchmark (section "Benchmarks"). *Bottom* Optical flow estimate based on MAP inference. The disk on the right displays the color code of flow vectors. Each image patch localized at x^i where a sufficiently discriminative feature could be extracted is associated with a set $\mathcal{U}(x^i)$ of possible assignment vectors $u_{\ell_i} \in \mathcal{U}(x^i)$. The displayed assignment field $\overline{u} := \{u_{\overline{\ell}_i}\}_{i \in V}$ is determined by a labeling field $\overline{\ell}$ minimizing the functional (115). The latter combinatorial task has been solved to global optimality by an approach combining convex relaxation and integer programming [93]. Global optimality enables model validation: any deficiencies of the assignment field estimate are solely due to the model components, feature extraction, and constraints, as encoded by the MRF (118) through $J(\ell; \mathcal{U})$

defined over the so-called *marginal polytope* $\mathcal{M}_G$. This polytope is the convex hull of feasible vectors μ, as is the simplex Δ_{m_i} in (120) for the feasible vectors μ^i. The combinatorial complexity of determining the integer-valued minimizer of (115) is reflected by the complexity of the marginal polytope $\mathcal{M}_G$. This complexity is due to the interaction of variables ℓ_i, ℓ_j as defined by the edges $ij \in E$ of the underlying graph, besides the integer constraints $\ell_i \in [m_i], \ \forall i \in [n]$.

Formulation (121) is the starting point for *convex relaxations* by optimizing over simpler polytopes, defined by a *subset* of inequalities that specify facets of $\mathcal{M}_G$. The recent paper [92] reports a comprehensive evaluation of a broad range of approaches to problem (121). Figure 15 illustrates an application to optical flow estimation.

While research on inference methods for graphical models is supporting the design of new approaches to optical flow estimation, the need to restrict the range of u to a finite set $\mathcal{U}$ is a significant restriction. As a consequence, approaches either exploit prior knowledge about u, so as to enable a covering of the relevant range of u with high resolution through the set $\mathcal{U}$ with bounded size $|\mathcal{U}|$, or solve problem (117) once more after refining $\mathcal{U}$, based on a first estimate of u.

For instance, the work [94] exploits the probabilistic model (118) in order to estimate locally the uncertainty of a first estimate u, which in turn is used to refine the set $\mathcal{U}$ so as to accommodate the discretization to the local variability of $u(x)$. The approach [95] first determines a coarse estimate of u in a preprocessing stage by global phase-based correlation, followed by defining possible refinements of $u(x)$ in terms of $\mathcal{U}$. The authors of [96]

rely on a prior estimate of the fundamental matrix F (34) using standard methods, which enables to properly define $\mathcal{U}$ based on the epipolar constraint (38).

In a way, while the former two approaches mimic range refinement of variational methods through representing u at multiple scales (section "Multiscale"), the latter approach exploits geometrical prior knowledge in a similar way to variational methods as discussed in section "Geometrical Prior Knowledge." Future research during the next decade will have to reveal more clearly the pros and cons of these related methods.

Variational Image Registration

The objective of *image registration* is to assign two images in terms of a diffeomorphism $u: \Omega \to \Omega$ of the underlying domain. A major motivation for this inherent *smoothness* of u has been applications to computational anatomy [97], based on fundamental work of Grenander, Dupuis, Trouvé, Miller, Younes, and others – cf. [97–100] and references therein.

Another basic motivation for the methodology of image registration is the use of point features, so-called landmarks, for establishing sparse assignments, which need to be interpolated in a subsequent step to obtain a transform of the entire underlying domain. This is usually accomplished by kernel functions that span a corresponding Hilbert space of smooth functions with bounded point-evaluation functional [101, 102]. Interpolation with thin-plate splines is a well-known example, and extensions to approximating deformations are straightforward. See [103, 104] for corresponding overviews in connection with medical imaging.

The *large deformation diffeomorphic metric matching (LDDMM)* approach [99, 100, 105] that emerged from the works cited above has evolved over the years into a theoretical and computational framework for diffeomorphic image registration. In particular, the application to the assignment of point sets, in connection with kernel functions, leads to a canonical system of ODEs whose numerical solution generates a diffeomorphic assignment along a geodesic path on the diffeomorphism group. See [106] for recent references and an extension for better handling deformations at multiple scales.

The importance of this framework is due to the well-developed mathematical basis and due to its broad applicability in the fields of computational anatomy and medical imaging. The mathematical relations to continuum and fluid mechanics and the corresponding relevancy to imaging problems with physical prior knowledge (cf. section "Physical Prior Knowledge") are intriguing as well. In the field of computer vision, deformable shape matching constitutes a natural class of applications, unlike the more common optical flow fields in natural videos that typically exhibit discontinuities, caused by depth changes and independently moving objects.

5 Open Problems and Perspectives

Unifying Aspects: Assignment by Optimal Transport

The mathematical theory of *optimal transport* [107, 108] provides a general formulation of the assignment problem that bears many relations to the approaches discussed so far.

Consider again the setup discussed in section "Assignment by Displacement Labeling": At each location x^i indexed by vertices $i \in V = [n]$, a vector $u(x^i) \in \mathcal{U}(x^i)$ from a set of candidates $\mathcal{U}(x^i)$ has to be selected. Put $\mathcal{U} = \cup_{i\in[n]}\mathcal{U}(x^i)$. Denote by V' the index set of all locations $\{x^i + u(x^i)\}_{u(x^i)\in\mathcal{U}(x^i)}$, $\forall i \in V$, that u may assign to the locations indexed by V. Then this setup is represented by the bipartite graph $G = (V, V'; E)$ with edge set $E = \{ij \in V \times V' : \exists u \in \mathcal{U},\ x^i + u = x^j\}$. The first term of the objective (115) specifies edge weights $\varphi_i(\ell_i; \mathcal{U})$ for each edge corresponding to the assignment $x^i + u_{\ell_i} = x^j$, and minimizing only the first term $\sum_{i\in V} \varphi_i(\ell_i; \mathcal{U})$ would *independently* select a unique vector $u(x^i)$ from each set $\mathcal{U}(x^i)$, $i \in V$, as solution to (120).

A classical way to remove this independency is to require the selection of *non-incident* assignments, that is, besides uniquely assigning a vector $u \in \mathcal{U}(x^i)$ to x^i, $\forall i \in V$, it is required that there is *at most one* correspondence $x^i + u = x^j$ for all $j \in V'$. This amounts to determining an optimal weighted *matching* in the bipartite graph $G = (V, V'; E)$. Formally, collecting the edge weights $\varphi_i(\ell_i; \mathcal{U})$ by a vector $\theta \in \mathbb{R}^m$, $m = \sum_{i\in V} m_i$, with subvectors given by (119), the LP

$$\min_{\mu\in\mathbb{R}^{|E(G)|}} \langle \theta, \mu \rangle \quad \text{subject to} \quad \mu \geq 0, \quad B_G\mu \leq \mathbb{1}_{|V\cup V'|}, \quad B_G \in \{0,1\}^{|V\cup V'|\times|E|}, \tag{122}$$

has to be solved where B_G is the incidence matrix of graph G. It is well known that the polyhedron $\mathbb{R}_+^{|E|} \cap \{\mu : B_G\mu \leq \mathbb{1}_{|V\cup V'|}\}$ is integral [109], which implies a binary solution $\overline{\mu} \in \{0,1\}^{|E|}$ to (122) satisfying the required uniqueness condition. Note that this condition may be regarded as a weak regularity condition enforcing a minimal degree of "smoothness" of the assignment field u.

The connection to optimal transport can be seen by reformulating problem (122). Put $n' = |V'|$ and let the matrix $c \in \mathbb{R}^{n\times n'}$ encode the *costs* of assigning (transporting) location x^i to $x^j = x^i + u$, $u \in \mathcal{U}(x^i)$. Then consider the problem

$$\min_{\mu\in\mathbb{R}^{n\times n'}} \langle c, \mu \rangle \text{ subject to } \mu \geq 0,\ \mu\mathbb{1}_{n'} = \mathbb{1}_n,\ \mu^\top\mathbb{1}_n \leq \mathbb{1}_{n'},\ n \leq n', \tag{123}$$

where the unknowns are deliberately denoted again by μ. The second constraint says that each node $i \in V$ (location x^i) is uniquely assigned to some node $j \in V'$ (location x^j). The third constraint says that at most one vertex $i \in V$ is assigned to each $j \in V'$. The last condition $n' \geq n$ naturally holds in practical applications. It is straightforward to show [110, Prop. 4.3] that the solution $\overline{\mu} \in \{0,1\}^{n\times n'}$ to (123) is again integral.

In the case $n = n'$, problem (123) equals the *linear assignment problem*, which is a discrete version of *Monge-Kantorovich* formulation of the *optimal transport* problem. The constraints of (123) then define the Birkhoff polytope, and the minimizer $\overline{\mu}$ at some vertex of this feasible set is a permutation matrix that uniquely maps V and V' onto each other. Matrices μ that are not vertices (extreme points) of the polytope are doubly stochastic; hence rows $\mu_{i,\bullet} \in \Delta_n$, $i \in [n]$ and columns $\mu_{\bullet,j}$, $j \in [n]$ represent *nondeterministic* assignments of vertices $i \in V$ and $j \in V'$, respectively.

The general formulation [107] considers Polish probability spaces $(\mathcal{X}, \mu_\mathcal{X})$, $(\mathcal{Y}, \mu_\mathcal{Y})$ with Borel probability measures $\mu_\mathcal{X} \in \mathcal{P}(\mathcal{X})$, $\mu_\mathcal{Y} \in \mathcal{P}(\mathcal{Y})$, and the set of *coupling measures*, again deliberately denoted by μ, that have $\mu_\mathcal{X}, \mu_\mathcal{Y}$ as marginals,

$$\mathcal{M}(\mu_\mathcal{X}, \mu_\mathcal{Y}) = \big\{\mu \in \mathcal{P}(\mathcal{X} \times \mathcal{Y}) : \mu(A \times \mathcal{Y}) = \mu_\mathcal{X}(A),\ \mu(\mathcal{X} \times B) = \mu_\mathcal{Y}(B),\ \forall A \subseteq \mathcal{B}(\mathcal{X}),\ \forall B \subseteq \mathcal{B}(\mathcal{Y})\big\}. \tag{124}$$

Given a Borel cost function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{+\infty\}$, the problem analogous to (123) in the case $n = n'$ reads

$$\inf_{\mu \in \mathcal{M}(\mu_\mathcal{X}, \mu_\mathcal{Y})} \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) \mathrm{d}\mu(x, y). \tag{125}$$

A central question concerns conditions on c that imply existence of *deterministic* minimizers $\overline{\mu}$ of (125), that is, existence of a measurable function $T : \mathcal{X} \to \mathcal{Y}$ such that for random variables (X, Y) with law μ, the relation $Y = T(X)$ holds. The *assignment* T is called *transportation map* that "pushes forward" the "mass" represented by $\mu_\mathcal{X}$ onto $\mu_\mathcal{Y}$, commonly denoted $T_\#$:

$$T_\# \mu_\mathcal{X} = \mu_\mathcal{Y} \qquad \text{with} \qquad \mu_\mathcal{Y}(B) = \mu_\mathcal{X}(T^{-1}(B)), \quad \forall B \in \mathcal{B}(\mathcal{Y}). \tag{126}$$

Likewise, $\overline{\mu}$ is concentrated on the graph of T, akin to the concentration of minimizers of (123) on a set of binary matrices.

Due to its generality formulation, (125) provides a single framework for addressing a range of problems, related to optical flow estimation by assignment. This particularly includes:

- The representation of both discrete and continuous settings, as sketched above, and the applicability to the assignment of arbitrary objects, as defined by the spaces $\mathcal{X}, \mathcal{Y}$.
- The focus on the combinatorial nature of the assignment problem, on convex duality and tightness or lack of tightness of the convex relaxation (125), together with a probabilistic interpretation in the latter case.
- Conservation of mass reflects the invariance assumption underlying (2) and (6), respectively.
- The differential, dynamic viewpoint: Let $\mathcal{X} = \mathbb{R}^d$ and define the cost function

$$c(x, y) = \|x - y\|^2 \tag{127}$$

and the *Wasserstein space* $\big(\mathcal{P}_2(\mathcal{X}), W_2\big)$ of Borel probability measures

$$\mathcal{P}_2(\mathcal{X}) := \big\{\mu_{\mathcal{X}} \in \mathcal{P}(\mathcal{X}) \colon \int_{\mathcal{X}} \|x - y\|^2 \mathrm{d}\mu_{\mathcal{X}}(x) < \infty,\ \forall y \in \mathcal{X}\big\}, \tag{128}$$

equipped with the *Wasserstein distance*

$$W_2(\mu_{\mathcal{X}}, \mu'_{\mathcal{X}}) := \left(\int_{\mathcal{X}\times\mathcal{X}} \|x - y\|^2 \mathrm{d}\overline{\mu}(x, y)\right), \qquad \forall \overline{\mu} \text{ solving (125)}, \tag{129}$$

with $\mu_{\mathcal{Y}}$ replaced by $\mu'_{\mathcal{X}}$ in (125). Then the path $(\mu_{\mathcal{X},t})$ defined by

$$\mu_{\mathcal{X},t} = \big((1-t)I + tT\big)_{\#}\mu_{\mathcal{X}} \tag{130}$$

and some optimal map T via (126) satisfies the continuity equation

$$\frac{\mathrm{d}}{\mathrm{d}t}\mu_t + \mathrm{div}(v_t\mu_t) = 0 \tag{131}$$

with velocity field $v_t \colon \mathbb{R}^d \to \mathbb{R}^d$ given by $v_t = (T - I) \circ ((1-t)I + tT)^{-1}$, $\forall t$ in the sense of distributions. Equation (131) provides a natural connection to continuum and fluid mechanics (cf., e.g., [111]) and also to flows generating diffeomorphic assignments under additional assumptions [100, Ch. 11]. Comparing (131) and (58) shows that, if g is regarded as a density for the scalar case $p = 1$, condition (57) is a strong assumption implying $\mathrm{div}u = 0$.

The generality of this framework explains too, however, why the regularity of solutions to the Monge-Kantorovich problem is a subtle issue, even when given as deterministic assignment T. This is also apparent through Euler's equation (131), which lacks any viscous term that would induce some regularity.

From this viewpoint, much of the research related to variational optical flow estimation, and to the related problems discussed in Sect. 4, can be understood as:

(i) Interplay between modeling additional terms that induce a desired degree of spatial regularity
(ii) Investigation of how this affects relaxation of the assignment problem from the optimization point of view and the accuracy of its solution.

As a consequence, *no sharp boundaries can* (and should) *be defined that separate these subfields of research*. For instance,

- The paper [112] suggested an early heuristic attempt to combine bipartite graph matching and thin-plate spline-based registration.
- The work [113] combines smoothing with radial basis functions and MRF-based labeling (section "Assignment by Displacement Labeling") for medical image registration.
- More generally, concerning image labeling, modeling spatial context by the edge-indexed terms φ_{ij} of the objective (115) entails the need to relax combinatorially complex polyhedral feasible sets like the marginal polytope in (121), whose vertices may *not* correspond to deterministic assignments, unlike assignments as solutions in the simpler case (122).
- The authors of [114] introduce a smoothing operator to solve numerically the Monge-Kantorovich problem.
- In [115] a related objective from continuum mechanics is proposed that, for a limiting value of some parameter, models a *viscous* fluid, hence ensures spatial regularity in a physically plausible way, as opposed to the pure continuity equation (131) that is lacking any such term. Assignments are computed by numerically tracing corresponding geodesic paths.
- Much more general objectives for assignments are addressed in [116] that take explicitly into account the metric structure of the underlying space $\mathcal{X}$. The problem to "linearize" this combinatorially complex objective in terms of the Monge-Kantorovich problem is studied in [110], along with the problem to define a cost function c so as to preserve the discriminative power of the original objective as much as possible.
- The recent work [117] exploits the Wasserstein distance (129) so as to solve simultaneously template-based assignment and image segmentation, by globally minimizing a corresponding joint variational objective.

This sample of the literature suggests to conclude that in the field of variational image registration (e.g., [100, 115]), sophisticated variational approaches exist that are satisfying in both respects (i),(ii) discussed above: these approaches clearly exhibit their properties mathematically, and they induce regularity without compromising accuracy of assignments, due to a good agreement with the physical properties of the objects being matched.

Outside these fields, a similar quality only holds for variational approaches to optical flow estimation that are constrained by – again: physically motivated – state equations (section "Physical Prior Knowledge"). A similar level of rigor has not been reached yet in a major application area of optical flow estimation: motion-based analysis of videos of unrestricted scenes with uncontrolled viewpoint changes and with independently moving rigid and articulated objects. This deficiency of related models is aggravated by the need for natural extensions of frame-to-frame assignments to the permanent analysis of *dynamic* scenarios over time (cf. section "Probabilistic Modeling and Online Estimation").

Motion Segmentation, Compressive Sensing

Research on *compressive sensing* [118, 119] and corresponding applications has been pervading all fields of empirical data analysis, including image reconstruction and more recently video analysis. A central theme is provable guarantees of signal recovery in polynomial runtime using sub-Nyquist sampling rates and convex relaxations of combinatorial objective functions for signal reconstruction. For instance, the most common scenario concerns the recovery of $u \in \mathbb{R}^n$ from $m \ll n$ linear measurements $Au = b \in \mathbb{R}^m$, by minimizing

$$\min_u \|u\|_1 \qquad \text{subject to} \qquad Au = b, \tag{132}$$

under the assumption that u is k-sparse, i.e.,

$$\|u\|_0 := |\mathrm{supp}(u)| = \big|\{i \in [n] : u_i \neq 0\}\big| \leq k. \tag{133}$$

The objective in (132) constitutes a convex relaxation of the combinatorial objective $\|u\|_0$, and suitable conditions on A, e.g., A is close to an isometry on the subset of $2k$-sparse vectors, guarantee unique recovery of u with high probability.

This section presents next an extension of this basic reconstruction principle to video analysis by sketching the recent work reported by [120]. Let

$$f_t \in \mathbb{R}^n, \qquad t \in [T], \tag{134}$$

denote the raw image sequence data in terms of vectorized image frames f_t, $t = 1, 2, \ldots, T$. Assuming a stationary camera as in surveillance applications, the objective is to separate the static background from objects moving in the foreground. The ansatz is based on the following modeling assumptions:

- At each point of time $t \in T$, image data are only sampled on a subset $\Omega_t \subset \Omega$ of the discretized domain Ω, resulting in subvectors

$$f_{\Omega_t}, \qquad t \in [T]. \tag{135}$$

  The sample set Ω_t may vary with t.
- The variation of components of f_{Ω_t} corresponding to the static background is caused by global smooth illumination changes. Hence, this part of f_{Ω_t} can be represented by a low-dimensional subspace

$$U_{\Omega_t} v_t, \qquad U_{\Omega_t} \in \mathbb{R}^{|\Omega_t| \times n_U}, \qquad t \in [T], \tag{136}$$

  generated by n_U orthonormal columns of a matrix U_t that are subsampled on Ω_t and some coefficient vector v_t. Research in computer vision [121, 122] supports this subspace assumption.

- Objects moving in the foreground cover only small regions within Ω. Hence they can be represented by vectors

$$s_{\Omega_t} \qquad \text{with} \qquad |\text{supp}(s)| \ll n. \tag{137}$$

Putting all together the model reads

$$f_{\Omega_t} = U_{\Omega_t} v_t + s_{\Omega_t}, \qquad t \in [T], \tag{138}$$

and convex relaxation of minimizing $|\text{supp}(s)|$ due to (137) leads to the recovery approach

$$\min_{U, v_t, s_{\Omega_t}} \|s_{\Omega_t}\|_1 \qquad \text{subject to} \qquad U_{\Omega_t} v_t + s_{\Omega_t} = f_{\Omega_t}. \tag{139}$$

Comparison to (132) shows similar usage of the sparsity-inducing ℓ^1 norm and subsampled measurements (135) as input data. On the other hand, the low-dimensional representation (136) of the static part of the video is estimated as well, and the entire video is recovered in terms of U_t (hence U rather than U_{Ω_t} is optimized in (139)). In fact, this joint optimization problem is *non*-convex and handled in [120] by alternating optimization:

– For *fixed* U_t, problem (139) is solved by applying ADMM (cf. section "Non-smooth Convex Functionals") to the augmented Lagrangian $L_\lambda(U, v_t, s_{\Omega_t}, w_{\Omega_t})$ with multiplier vector w_{Ω_t} and parameter λ as in (95).
– Having determined $v_t, s_{\Omega_t}, w_{\Omega_t}$, the subspace U_t is tracked by performing gradient descent with respect to $L(\cdot, v_t, s_{\Omega_t}, w_{\Omega_t})$ on the Grassmannian $\mathcal{G}(n_U, \mathbb{R}^n)$ (cf., e.g., [123]), resulting in U_{t+1}.

The closely related static viewpoint on the same problem reveals its relevancy to several important research directions. Let

$$F = [f_1, \ldots, f_T] = L + S \tag{140}$$

denote the whole video data that, due to the reasoning above, are supposed to be decomposable into a *low-rank* matrix L and a *sparse* matrix S. The corresponding convex relaxation approach [124] reads

$$\min_{L,S} \|L\|_* + \alpha \|S\|_1 \qquad \text{subject to} \qquad L + S = F, \tag{141}$$

where $\|L\|_* = \sum_i \sigma_i(L)$ denotes the nuclear norm in terms of the singular values of L and $\|S\|_1 = \sum_{i,j} |S_{ij}|$. Here, the nuclear norm $\|\cdot\|_*$ constitutes a convex relaxation of the combinatorial task to minimize the rank of L, analogous to replacing the combinatorial objective $\|u\|_0$ in (133) by $\|u\|_1$ in (132). Clearly, the online ansatz (138) along with the corresponding incremental estimation approach

is more natural for processing *long* videos. The price to pay is the need to cope with a non-convex (albeit smooth) problem, whereas the batch approach (141) is convex.

Future research will tackle the challenging, more general case of non-static backgrounds and moving cameras, respectively. For scenarios with small displacements $u(x)$, work that represents the state of the art is reported in [125]. Results in computer vision that support subspace models and low-rank assumptions have been established [126], and the problem of clustering data lying in unknown low-dimensional subspaces has received considerable attention [127–129].

From a broader perspective, video analysis and motion-based segmentation provide attractive connections to research devoted to union-of-subspaces models of empirical data and relevant compressive sensing principles [130–132] and to advanced *probabilistic* models and methods for nonparametric inference [133, 134].

Probabilistic Modeling and Online Estimation

There is a need for advanced probabilistic models, and three related aspects of increasing difficulty are briefly addressed:

- A persistent issue of most variational models of mathematical imaging, including those for optical flow estimation, concerns the *selection of* appropriate *hyperparameter values*, like the parameter σ of (62) weighting the combination of data term and regularizer (56). In principle, Bayesian hierarchical modeling [135] provides the proper framework for calibrating variational models in this respect. The paper [136] illustrates an application in connection with optical flow estimation, based on the marginal data likelihood [137] interpreted as hyperparameter (model) evidence.

  Estimating hyperparameter values from the given data in this way entails the evaluation of high-dimensional integrals for marginalization, commonly done using Laplace's method and a corresponding approximation by Gaussian (quadratic) integrals [138, 139]. A validation for complex high-dimensional posterior distributions encountered in variational imaging is involved, however, and is also stimulating more recent research in the field of statistics [140].

  Using discrete variational models (section "Assignment by Displacement Labeling") aggravates this problem, due to considerable computational costs and since no widely accepted methods have been established analogous to the abovementioned approximations.
- Computational costs in connection with runtime requirements become a serious problem when *dynamic scenarios* are considered. While extensions of the domain to $\Omega \times [0, T]$ like in (82) are straightforward mathematically and have proven to significantly increase accuracy of optical flow estimation, employing a static model in terms of elliptic Euler-Lagrange systems to a dynamic system appears somewhat odd, not to mention the need to shift the time interval $[0, T]$ along the time axis in order to analyze long image sequences.

Such extensions appear more natural in connection with dynamic physical models constraining optical flow estimation, as opposed to stationary formulations like (88). See [141] for a corresponding approach to data assimilation [142]. A nice feature of this method is the ability to estimate initial conditions that are generally unknown, too. On the other hand, the computational costs necessitate to propagate a low-dimensional POD-projection of the state variables (POD: proper orthogonal decomposition) since the control of dynamical systems [70] entails looping forward and backward through the entire time interval.

- The last remark points to the need for *online estimation methods* that are *causal and optimal*, in connection with the analysis of dynamical system through image analysis. Again the proper framework is known since decades: Given stochastic state and observation processes

$$S = \{S_t\}_{t\geq 0}, \qquad G = \{G_t\}_{t\geq 0}, \tag{142}$$

stochastic filtering[143] amounts to determine the conditional distribution of S_t given the observation history and to evaluate it in terms of expectations of the form $\mathbb{E}[\varphi(S_t)|g_s,\, 0 \leq s \leq t]$, for some statistic $\varphi(\cdot)$ of interest (e.g., simply $\varphi(S_t) = S_t$) and conditioned on realizations g_s of G_s, $s \in [0, t]$. Most research during the last decade considered the design of particle filters [143, 144] to the estimation of *low*-dimensional states based on image measurements. This does not scale-up however to high-dimensional states like optical flows $S_t = u_t$.

An attempt to mimic online estimation in connection with instationary optical flows related to experimental fluid dynamics is presented in [68], with states and their evolution given by vorticity transport. For low signal-to-noise ratios and sufficiently high frame rates, the approach performs remarkably well. Another dynamical computer vision scenario is discussed in the recent work [57]. Here the states $S_t = (z_t, \{h_t, R_t\}) \in \mathbb{R}^n \times \mathsf{SE}(3)$ are dense depth-maps z_t (cf. (83)) together with varying motion parameters $\{h_t, R_t\}$ describing the observer's motion relative to the scene, to be estimated from image sequence features g_t as measurements via optical flow estimates u_t – see Fig. 13. The approach involves prediction and fusion steps based on Gaussian approximation and joint optimization, yet cannot be considered as direct application of the stochastic filtering framework, in a strict sense. This assessment applies also to labeling approaches (section "Assignment by Displacement Labeling") and their application to dynamic scenarios.

6 Conclusion

Optical flow estimates form an essential basis for low-level and high-level image sequence analysis and thus are relevant to a wide range of applications. Corresponding key problems, concepts, and their relationships were presented, along with numerous references to the literature for further study. Despite three decades of research, however, an overall coherent framework that enables to mathematically

model, predict, and estimate the performance of corresponding computational systems in general scenarios is still lacking. This short survey will hopefully stimulate corresponding methodological research.

7 Basic Notation

List of major symbols used in the text

Symbol	Brief description	Reference
r.h.s.	abbr.: right-hand side (of some equation)	
w.r.t.	abbr.: with respect to	
w.l.o.g.	without loss of generality	
LP	linear program	
$\mathbb{1}_n \in \mathbb{R}^n$	$(1, 1, \ldots, 1)^\top$	
$[n],\ n \in \mathbb{N}$	integer range $\{1, 2, \ldots, n\}$	
$[n]_0,\ n \in \mathbb{N}$	integer range $\{0, 1, \ldots, n-1\}$	
$\Omega \subset \mathbb{R}^d$	image domain; typically $d \in \{2, 3\}$	
$x = \left(x_1, \ldots, x_d\right)^\top \in \Omega$	image point	
$u(x,t) \in \mathbb{R}^d$	assignment, motion or optical flow field	(4), (7), (8) Sections "Assignment Fields" and "Motion Fields"
$X = (X_1, X_2, X_3)^\top \in \mathbb{R}^3$	scene point	
$y \in \mathbb{P}^2,\ Y \in \mathbb{P}^3$	homogeneous representation of image and scene points x and X, resp.	Section "Two-View Geometry"
$\mathsf{SO}(3),\ \mathfrak{so}(3)$	special orthog. group and its Lie algebra	
$\mathsf{SE}(3)$	group of Euclidean (rigid) transf. of $\mathbb{R}^3$	
$\{h, R\} \in \mathsf{SE}(3)$	Euclidean transformation of $\mathbb{R}^3$	(28)
$[q]_\times \in \mathfrak{so}(3),\ q \in \mathbb{R}^3$	skew-symm. matrix defined by	

(continued)

Symbol	Brief description	Reference
	$[q]_\times X = q \times X,\ \forall X \in \mathbb{R}^3$	
$K \in \mathbb{R}^{3\times 3}$	camera matrix (internal parameters)	Section "Two-View Geometry"
$F, E \in \mathbb{R}^{3\times 3}$	fundamental and essential matrix	Section "Two-View Geometry"
$f(x,t),\ x \in \Omega,\ t \in \mathbb{R}$	image sequence	
$\partial_i = \frac{\partial}{\partial x_i},\ i \in [d]$	spatial partial derivative	
$\partial_t = \frac{\partial}{\partial t}$	temporal partial derivative	
$\partial^\alpha = \frac{\partial^{\lvert\alpha\rvert}}{\partial_1^{\alpha_1}\cdots\partial_d^{\alpha_d}}$ $\alpha \in \mathbb{N}^d$, $\lvert\alpha\rvert = \sum_{i\in[d]} \alpha_i$	multi-index notation	
$\omega^\alpha = \omega_1^{\alpha_1}\cdots\omega_d^{\alpha_d}$	monomial from $\omega \in \mathbb{R}^d$	
$\nabla f(x,t) = \begin{pmatrix} \partial_1 f(x,t) \\ \cdots \\ \partial_d f(x,t) \end{pmatrix}$	spatial gradient	
$\nabla_t f(x,t) = \begin{pmatrix} \nabla f(x,t) \\ \partial_t f(x,t) \end{pmatrix}$	spatio-temporal gradient	
$\mathrm{div} u$	divergence $\sum_{i\in[d]} \partial_i u_i$ of a vector field u	
Δ	Laplace operator $\sum_{i\in[d]} \partial_i^2$	
$g(x,t) \in \mathbb{R}^p,\ p \geq 1$	feature mapping (specific meaning and p depend on the context)	
$\mathrm{J}_g(x) = \left(\left(\nabla g_i(x)\right)_j\right)_{i\in[p], j\in[d]}$	Jacobian matrix of $g(x) \in \mathbb{R}^p$ at $x \in \mathbb{R}^d$	
$\mathrm{J}_{g,t}(x,t) = \left(\left(\nabla_t g_i(x,t)\right)_j\right)_{\substack{i\in[p] \\ j\in[d]\cup\{t\}}}$	Jacobian of $g(x,t) \in \mathbb{R}^p$ at $(x,t) \in \mathbb{R}^{d+1}$	
$\hat{g}(\omega) = \mathcal{F}g(\omega) = \big(\mathcal{F}g\big)(\omega)$	Fourier transform of g	page 1952, (17)
$\langle x, x'\rangle = \sum_i x_i x_i'$	Euclidean inner product	
$\lVert x\rVert = \langle x, x\rangle^{1/2}$	Euclidean ℓ^2 norm	
$\lVert x\rVert_1 = \sum_i \lvert x_i\rvert$	ℓ^1 norm	
$\mathrm{diag}(x)$	diagonal matrix with vector x as diagonal	
$\ker A$	nullspace of the linear mapping A	
$\mathrm{tr} A = \sum_i A_{i,i}$	trace of matrix A	

(continued)

Symbol	Brief description	Reference
$\langle A, B\rangle = \mathrm{tr}(A^\top B)$	matrix inner product	
$\|A\|_F = \langle A, A\rangle^{1/2}$	Frobenius norm	
$\rho\colon \mathbb{R} \to \mathbb{R}_+$	distance function	page 1948, (3)
$\delta_C(x) = \begin{cases} 0, & x \in C \\ +\infty, & x \notin C \end{cases}$	indicator function of a closed convex set $C \subseteq \mathbb{R}^d$	
Π_C	orthogonal projection onto a closed convex set C	
$\Delta_n \subset \mathbb{R}^n$	probability simplex $\{x \in \mathbb{R}^n\colon \sum_{i\in[n]} x_i = 1;\ x \geq 0\}$	page 1985, (119)

Cross-References

- Compressive Sensing
- Duality and Convex Programming
- Energy Minimization Methods
- Iterative Solution Methods
- Large-Scale Inverse Problems in Imaging
- Linear Inverse Problems
- Non-Linear Image Registration
- Regularization Methods for Ill-Posed Problems
- Splines and Multiresolution Analysis
- Statistical Methods in Imaging
- Total Variation in Imaging

References

1. Margarey, J., Kingsbury, N.: Motion estimation using a complex-valued wavelet transform. IEEE Trans. Signal Process. **46**(4), 1069–1084 (1998)
2. Bernard, C.: Discrete wavelet analysis for fast optic flow computation. Appl. Comput. Harmon. Anal. **11**, 32–63 (2001)
3. Lowe, D.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (2004)
4. Bay, H., Ess, A., Tuytelaars, T., Ban Gool, L.: Speeded-up robust features (SURF). Comput. Vis. Image Underst. **110**(3), 346–359 (2008)
5. Brown, M., Burschka, D., Hager, G.: Advances in computational stereo. IEEE Trans. Pattern Anal. Mach. Intell. **25**(8), 993–1008 (2003)
6. Morel, J.M., Yu, G.: ASIFT: a new framework for fully affine invariant image comparison. SIAM J. Imaging Sci. **2**(2), 438–469 (2009)

7. Sabater, S., Almansa, A., Morel, J.: Meaningful matches in stereovision. IEEE Trans. Pattern Anal. Mach. Intell. **34**(5), 930–942 (2012)
8. Black, M., Anandan, P.: The robust estimation of multiple motions: parametric and piecewise-smooth flow fields. Comput. Vis. Image Underst. **63**(1), 75–104 (1996)
9. Auslender, A., Teboulle, M.: Asymptotic Cones and Functions in Optimization and Variational Inequalities. Springer, New York (2003)
10. Rockafellar, R., Wets, R.J.B.: Variational Analysis, 2nd edn. Springer, Berlin/New York (2009)
11. Sun, D., Roth, S., Black, M.: A quantitative analysis of current practices in optical flow estimation and the principles behind them. Int. J. Comput. Vis. **106**(2), 115–137 (2013)
12. Bruhn, A., Weickert, J., Kohlberger, T., Schnörr, C.: A multigrid platform for real-time motion computation with discontinuity-preserving variational methods. Int. J. Comput. Vis. **70**(3), 257–277 (2006)
13. Gwosdek, P., Bruhn, A., Weickert, J.: Variational optic flow on the Sony Playstation 3 – accurate dense flow fields for real-time applications. J. Real-Time Image Process. **5**(3),163–177 (2010)
14. Hartley, R., Zisserman, A.: Multiple View Geometry in Computer Vision. Cambridge University Press, Cambridge (2000)
15. Faugeras, O., Luong, Q.T.: The Geometry of Multiple Images. MIT, Cambridge/London (2001)
16. Longuet-Higgins, H., Prazdny, K.: The interpretation of a moving retinal image. Proc. R. Soc. Lond. B **208**, 385–397 (1980)
17. Prazdny, K.: Egomotion and relative depth map from optical flow. Biol. Cybern. **36**, 87–102 (1980)
18. Kanatani, K.: Transformation of optical flow by camera rotation. IEEE Trans. Pattern Anal. Mach. Intell. **10**(2), 131–143 (1988)
19. Lucas, B., Kanade, T.: An iterative image registration technique with an application to stereo vision. In: Proceedings of the IJCAI, Vancouver, vol. 2, pp. 674–679 (1981)
20. Nagel, H.H.: Constraints for the estimation of displacement vector fields from image sequences. In: Proceedings of the International Joint Conference on Artificial Intelligence, Karlsruhe, pp. 945–951 (1983)
21. Hildreth, E.: The computation of the velocity field. Proc. R. Soc. B **221**, 189–220 (1984)
22. Werkhoven, P., Toet, A., Koenderink, J.: Displacement estimates through adaptive affinities. IEEE Trans. Pattern Anal. Mach. Intell. **12**(7), 658–663 (1990)
23. Werkhoven, P., Koenderink, J.: Extraction of motion parallax structure in the visual system I. Biol. Cybern. **83**, 185–191 (1990)
24. Werkhoven, P., Koenderink, J.: Extraction of motion parallax structure in the visual system II. Biol. Cybern. **63**, 193–199 (1990)
25. Verri, A., Poggio, T.: Motion field and optical flow: qualitative properties. IEEE Trans. Pattern Anal. Mach. Intell. **11**(5), 490–498 (1989)
26. Adelson, E., Bergen, J.: Spatiotemporal energy models for the perception of motion. J. Opt. Soc. Am. A **2**(2), 284–299 (1985)
27. Heeger, D.: Optical flow using spatiotemporal filters. Int. J. Comput. Vis. **1**(4), 279–302 (1988)
28. Fleet, D., Jepson, A.: Computation of component image velocity from local phase information. Int. J. Comput. Vis. **5**(1), 77–104 (1990)
29. Horn, B., Schunck, B.: Determining optical flow. Artif. Intell. **17**(1–3), 185–203 (1981)
30. Nagel, H.H., Enkelmann, W.: An investigation of smoothness constraints for the estimation of displacement vector fields from image sequences. IEEE Trans. Pattern Anal. Mach. Intell. **8**(5), 565–593 (1986)
31. Anandan, P.: A computational framework and an algorithm for the measurement of visual motion. Int. J. Comput. Vis. **2**, 283–310 (1989)
32. Yuille, A., Grzywacz, N.: A mathematical analysis of the motion coherence theory. Int. J. Comput. Vis. **3**, 155–175 (1989)

33. Schnörr, C.: Determining optical flow for irregular domains by minimizing quadratic functionals of a certain class. Int. J. Comput. Vis. **6**(1), 25–38 (1991)
34. Hinterberger, W., Scherzer, O., Schnörr, C., Weickert, J.: Analysis of optical flow models in the framework of calculus of variations. Numer. Funct. Anal. Optim. **23**(1/2), 69–89 (2002)
35. Weickert, J., Schnörr, C.: Variational optic flow computation with a spatio-temporal smoothness constraint. J. Math. Imaging Vis. **14**(3), 245–255 (2001)
36. Weickert, J., Schnörr, C.: A theoretical framework for convex regularizers in PDE–based computation of image motion. Int. J. Comput. Vis. **45**(3), 245–264 (2001)
37. Wang, J., Adelson, E.: Representing moving images with layers. IEEE Trans. Image Process. **3**(5), 625–638 (1994)
38. Tomasi, C., Kanade, T.: Shape and motion from image streams under orthography: a factorization method. Int. J. Comput. Vis. **9**(2), 137–154 (1992)
39. Schnörr, C.: Computation of discontinuous optical flow by domain decomposition and shape optimization. Int. J. Comput. Vis. **8**(2), 153–165 (1992)
40. Heitz, F., Bouthemy, P.: Multimodal estimation of discontinuous optical flow using Markov random fields. IEEE Trans. Pattern Anal. Mach. Intell. **15**(12), 1217–1231 (1993)
41. Barron, J.L., Fleet, D., Beauchemin, S.S.: Performance of optical flow techniques. Int. J. Comput. Vis. **12**(1), 43–77 (1994)
42. Baker, S., Scharstein, D., Lewis, J.P., Roth, S., Black, M.J., Szeliski, R.: A database and evaluation methodology for optical flow. Int. J. Comput. Vis. **92**(1), 1–31 (2011)
43. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: the KITTI dataset. Int. J. Robot. Res. (IJRR) **32**(11), 1231–1237 (2013)
44. Butler, D.J., Wulff, J., Stanley, G.B., Black, M.J.: A naturalistic open source movie for optical flow evaluation. In: Fitzgibbon, A., et al. (eds.) Proceedings of the ECCV, Part IV, Florence. LNCS vol. 7577, pp. 611–625. Springer (2012)
45. Saad, Y.: Iterative Methods for Sparse Linear Systems. SIAM, Philadelphia (2003)
46. Schnörr, C.: On functionals with greyvalue-controlled smoothness terms for determining optical flow. IEEE Trans. Pattern Anal. Mach. Intell. **15**(10), 1074–1079 (1993)
47. Bruhn, A., Weickert, J., Schnörr, C.: Lucas/Kanade meets Horn/Schunck: combining local and global optic flow methods. Int. J. Comput. Vis. **61**(3), 211–231 (2005)
48. Yuan, J., Schnörr, C., Mémin, E.: Discrete orthogonal decomposition and variational fluid flow estimation. J. Math. Image Vis. **28**, 67–80 (2007)
49. Yuan, J., Schnörr, C., Steidl, G.: Simultaneous optical flow estimation and decomposition. SIAM J. Sci. Comput. **29**(6), 2283–2304 (2007)
50. Yuan, J., Schnörr, C., Steidl, G.: Convex Hodge decomposition and regularization of image flows. J. Math. Imaging Vis. **33**(2), 169–177 (2009)
51. Werlberger, M., Trobin, W., Pock, T., Wedel, A., Cremers, D., Bischof, H.: Anisotropic Huber-L1 optical flow. In: Proceedings of the BMVC, London (2009)
52. Krähenbühl, P., Koltun, V.: Efficient nonlocal regularization for optical flow. In: Fitzgibbon, A.W., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) Proceedings of the ECCV, Florence, LNCS, vol. 7572, pp. 356–369. Springer (2012)
53. Kindermann, S., Osher, S., Jones, P.: Deblurring and denoising of images by nonlocal functionals. Multiscale Model. Simul. **4**(4), 1091–1115 (2005)
54. Gilboa, G., Osher, S.: Nonlocal operators with applications to image processing. Multiscale Model. Simul. **7**(3), 1005–1028 (2008)
55. Elmoataz, A., Lezoray, O., Bougleux, S.: Nonlocal discrete regularization on weighted graphs: a framework for image and manifold processing. IEEE Trans. Image Process. **17**(7), 1047–1059 (2008)
56. Volz, S., Bruhn, A., Valgaerts, L., Zimmer, H.: Modeling temporal coherence for optical flow. In: Metaxas, D.N., Quan, L., Sanfeliu, A., Gool, L.J.V. (eds.) Proceedings of the ICCV, Barcelona, pp. 1116–1123. IEEE (2011)
57. Becker, F., Lenzen, F., Kappes, J.H., Schnörr, C.: Variational recursive joint estimation of dense scene structure and camera motion from monocular high speed traffic sequences. Int. J. Comput. Vis. **105**, 269–297 (2013)

58. Slesareva, N., Bruhn, A., Weickert, J.: Optic flow goes stereo: a variational method for estimating discontinuity preserving dense disparity maps. In: Proceedings of the 27th DAGM Symposium, Vienna, pp. 33–40 (2005)
59. Valgaerts, L., Bruhn, A., Mainberger, M., Weickert, J.: Dense versus sparse approaches for estimating the fundamental matrix. Int. J. Comput. Vis. **96**(2), 212–234 (2012)
60. Adrian, R.J., Westerweel, J.: Particle Image Velocimetry. Cambridge University Press, Cambridge/New York (2011)
61. Arnaud, E., Mémin, E., Sosa, R., Artana, G.: A fluid motion estimator for Schlieren image velocimetry. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) Proceedings of the ECCV, Graz. LNCS, vol. 3951, pp. 198–210. Springer (2006)
62. Westerweel, J.: Fundamentals of digital particle image velocimetry. Meas. Sci. Technol. **8**, 1379–1392 (1998)
63. Ruhnau, P., Gütter, C., Putze, T., Schnörr, C.: A variational approach for particle tracking velocimetry. Meas. Sci. Technol. **16**, 1449–1458 (2005)
64. Ruhnau, P., Kohlberger, T., Nobach, H., Schnörr, C.: Variational optical flow estimation for particle image velocimetry. Exp. Fluids **38**, 21–32 (2005)
65. Heitz, D., Mémin, E., Schnörr, C.: Variational fluid flow measurements from image Sequences: synopsis and perspectives. Exp. Fluids **48**(3), 369–393 (2010)
66. Vlasenko, A., Schnörr, C.: Physically consistent and efficient variational denoising of image fluid flow estimates. IEEE Trans. Image Process. **19**(3), 586–595 (2010)
67. Ruhnau, P., Schnörr, C.: Optical Stokes flow estimation: an imaging-based control approach. Exp. Fluids **42**, 61–78 (2007)
68. Ruhnau, P., Stahl, A., Schnörr, C.: Variational estimation of experimental fluid flows with physics-based spatio-temporal regularization. Meas. Sci. Technol. **18**, 755–763 (2007)
69. Girault, V., Raviart, P.A.: Finite Element Methods for Navier-Stokes Equations. Springer, Berlin/New York (1986)
70. Gunzburger, M.: Perspectives in Flow Control and Optimization. SIAM, Philadelphia (2003)
71. Ito, K., Kunisch, K.: Lagrange Multiplier Approach to Variational Problems and Applications. Advances in Design and Control, vol. 15. SIAM, Philadelphia (2008)
72. Gunzburger, M., Manservisi, S.: Analysis and approximation of the velocity tracking problem for Navier-Stokes flows with distributed control. SIAM J. Numer. Anal. **37**(5), 1481–1512 (2000)
73. Chambolle, A., Pock, T.: A first-order primal-dual algorithm for convex problems with applications to imaging. J. Math. Imaging Vis. **40**(1), 120–145 (2011)
74. Briggs, W., Henson, V., McCormick, S.: A Multigrid Tutorial, 2nd edn. SIAM, Philadelphia (2000)
75. Bruhn, A., Weickert, J., Feddern, C., Kohlberger, T., Schnörr, C.: Variational optic flow computation in real-time. IEEE Trans. Image Process. **14**(5), 608–615 (2005)
76. Trottenberg, U., Oosterlee, C., Schüller, A.: Multigrid. Academic, San Diego (2001)
77. Bach, F., Jenatton, R., Mairal, J., Obozinski, G.: Optimization with sparsity-inducing penalties. Found. Trends Mach. Learn. **4**(1), 1–106 (2012)
78. Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J.: Distributed optimization and statistical learning via the alternating direction method of multipliers. Found. Trends Mach. Learn. **3**(1), 1–122 (2010)
79. Combettes, P., Pesquet, J.C.: Proximal splitting methods in signal Processing. In: Bauschke, H., Burachik, R., Combettes, P., Elser, V., Luke, D., Wolkowicz, H. (eds.) Fixed-Point Algorithms for Inverse Problems in Science and Engineering. Springer, New York (2010)
80. Parikh, N., Boyd, S.: Proximal algorithms. Found. Trends Optim. **1**(3), 1–108 (2013)
81. Horst, R., Thoai, N.: DC programming: overview. J. Optim. Theory Appl. **103**(1), 1–43 (1999)
82. Hoai An, L., Pham Dinh, T.: The DC (difference of convex functions) programming and DCA revisited with DC models of real world Nonconvex optimization Problems. Ann. Oper. Res. **133**, 23–46 (2005)

83. Steinbrücker, F., Pock, T., Cremers, D.: Advanced data terms for variational optic flow estimation. In: Magnor, M.A., Rosenhahn, B., Theisel, H. (eds.) Proceedings Vision, Modeling and Visualization (VMV), Braunschweig, pp. 155–164. DNB (2009)
84. Hafner, D., Demetz, O., Weickert, J.: Why is the census transform good for robust optic flow computation? In: Kuijper, A., Bredies, K., Pock, T., Bischof, H. (eds.) Proceedings of the SSVM, Leibnitz. LNCS, vol. 7893, pp. 210–221. Springer (2013)
85. Viola, P., Wells, W.M. III: Alignment by maximization of mutual information. Int. J. Comput. Vis. **24**(2), 137–154 (1997)
86. Hirschmuller, H.: Stereo processing by semiglobal matching and mutual information. IEEE Trans. Pattern Anal. Mach. Intell. **30**(2), 328–341 (2008)
87. Cover, T., Thomas, J.: Elements of Information Theory. Wiley, New York (1991)
88. Parzen, E.: On the estimation of a probability density function and the mode. Ann. Math. Stat. **33**, 1065–1076 (1962)
89. Kim, J., Kolmogorov, V., Zabih, R.: Visual correspondence using energy minimization and mutual information. In: Proceedings of the ICCV, Nice (2003)
90. Wainwright, M., Jordan, M.: Graphical models, exponential families, and variational inference. Found. Trends Mach. Learn. **1**(1–2), 1–305 (2008)
91. Koller, D., Friedman, N.: Probabilistic Graphical Models: Principles and Techniques. MIT, Cambridge (2009)
92. Kappes, J., Andres, B., Hamprecht, F., Schnörr, C., Nowozin, S., Batra, D., Kim, S., Kausler, B., Lellmann, J., Komodakis, N., Rother, C.: A comparative study of modern inference techniques for discrete energy minimization problem. In: Proceedings of the CVPR, Portland (2013)
93. Savchynskyy, B., Kappes, J., Swoboda, P., Schnörr, C.: Global MAP-optimality by shrinking the combinatorial search area with convex relaxation. In: Proceedings of the NIPS, Lake Tahoe (2013)
94. Glocker, B., Paragios, N., Komodakis, N., Tziritas, G., Navab, N.: Optical flow estimation with uncertainties through dynamic MRFs. In: Proceedings of the CVPR, Anchorage (2008)
95. Mozerov, M.: Constrained optical flow estimation as a matching problem. IEEE Trans. Image Process. **22**(5), 2044–2055 (2013)
96. Yamaguchi, K., McAllester, D., Urtasun, R.: Robust monocular epipolar flow estimation. In: Proceedings of the CVPR, Portland (2013)
97. Younes, L., Arrate, F., Miller, M.: Evolution equations in computational anatomy. NeuoImage **45**(1, Suppl. 1), S40–S50 (2009)
98. Dupuis, P., Grenander, U., Miller, M.: Variational problems on flows of diffeomorphisms for image matching. Q. Appl. Math. **56**(3), 587–600 (1998)
99. Beg, M., Miller, M., Trouv'e, A., Younes, L.: Computing large deformation metric mappings via geodesic flows of diffeomorphisms. Int. J. Comput. Vis. **61**(2), 139–157 (2005)
100. Younes, L.: Shapes and Diffeomorphisms. Applied Mathematical Sciences, vol. 171. Springer, Heidelberg/New York (2010)
101. Whaba, G.: Spline Models for Observational Data. SIAM, Philadelphia (1990)
102. Buhmann, M.: Radial Basis Functions. Cambridge University Press, Cambridge/New York (2003)
103. Rohr, K.: Landmark-Based Image Analysis. Kluwer Academic, Dordrecht/Boston (2001)
104. Modersitzki, J.: Numerical Methods for Image Registration. Oxford University Press, Oxford/New York (2004)
105. Glaunès, J., Qiu, A., Miller, M., Younes, L.: Large deformation diffeomorphic metric curve mapping. Int. J. Comput. Vis. **80**(3), 317–336 (2008)
106. Sommer, S., Lauze, F., Nielsen, M., Pennec, X.: Sparse multi-scale diffeomorphic registration: the kernel bundle framework. J. Math. Imaging Vis. **46**(3), 292–308 (2013)
107. Villani, C.: Optimal Transport: Old and New. Springer, Berlin (2009)
108. Ambrosio, L., Gigli, N.: A user's guide to optimal transport. In: Modeling and Optimisation of Flows on Networks, Cetraro Lecture Notes in Mathematics, vol. 2062, pp. 1–155. Springer (2013)

109. Korte, B., Vygen, J.: Combinatorial Optimization, 4th edn. Springer, Berlin (2008)
110. Schmitzer, B., Schnörr, C.: Modelling convex shape priors and matching based on the Gromov-Wasserstein distance. J. Math. Imaging Vis. **46**(1), 143–159 (2013)
111. Benamou, J.D., Brenier, Y.: A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. Numer. Math. **84**(3), 375–393 (2000)
112. Belongie, S., Malik, J., Puzicha, J.: Shape matching and object recognition using shape contexts. IEEE Trans. Pattern Analysis Mach. Intell. **24**(24), 509–522 (2002)
113. Glocker, B., Komodakis, N., Tziritas, G., Navab, N., Paragios, N.: Dense image registration through MRFs and efficient linear programming. Med. Image Anal. **12**, 731–741 (2008)
114. Angenent, S., Haker, S., Tannenbaum, A.: Minimizing flows for the Monge–Kantorovich problem. SIAM J. Math. Anal. **35**(1), 61–97 (2003)
115. Wirth, B., Bar, L., Rumpf, M., Sapiro, G.: A continuum mechanical approach to geodesics in shape space. Int. J. Comput. Vis. **93**(3), 293–318 (2011)
116. Mémoli, F.: Gromov-Wasserstein distances and the metric approach to object matching. Found. Comput. Math. **11**, 417–487 (2011)
117. Schmitzer, B., Schnörr, C.: Object segmentation by shape matching with Wasserstein Modes. In: Proceedings of the EMMCVPR, Lund. Springer (2013)
118. Candès, E., Romberg, J., Tao, T.: Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. IEEE Trans. Inf. Theory **52**, 489–509 (2006)
119. Donoho, D.: Compressed sensing. IEEE Trans. Inf. Theory **52**, 1289–1306 (2006)
120. He, J., Balzano, L., Szlam, A.: Incremental gradient on the Grassmannian for online foreground and background separation in subsampled video. In: Proceedings of the CVPR, Providence (2012)
121. Belhumeur, P., Kriegman, D.: What is the set of images of an object under all possible illumination conditions? Int. J. Comput. Vis. **28**(3), 245–260 (1998)
122. Basri, R., Jacobs, D.: Lambertian reflectance and linear subspaces. IEEE Trans. Pattern Anal. Mach. Intell. **25**(2), 218–233 (2003)
123. Absil, P.A., Mahony, R., Sepulchre, R.: Optimization Algorithms on Matrix Manifolds. Princeton University Press, Princeton/Woodstock (2008)
124. Candés, E., Li, X., Ma, Y., Wright, J.: Robust principal component analysis? J. ACM **58**(3), Article no. 11 (2011)
125. Ayvaci, A., Raptis, M., Soatto, S.: Sparse occlusion detection with optical flows. Int. J. Comput. Vis. **97**(3), 322–338 (2012)
126. Irani, M.: Multi-frame correspondence estimation using subspace constraints. Int. J. Comput. Vis. **48**(3), 173–194 (2002)
127. Liu, G., Lin, Z., Yan, S., Sun, J., Yu, Y., Ma, Y.: Robust recovery of subspace structures by low-rank representation. IEEE Trans. Pattern Anal. Mach. Intell. **35**(1), 171–184 (2013)
128. Aldroubi, A.: The subspace segmentation problem, nonlinear approximations and applications. ISRN Signal Proc. Art.(417492), 13p (2013)
129. Elhamifar, E., Vidal, R.: Sparse subspace clustering: algorithm, theory, and applications. IEEE Trans. Pattern Anal. Mach. Intell. **35**(11), 2765–2781 (2013)
130. Lu, Y., Do, M.: A theory for samping signals from a union of subspaces. IEEE Trans. Signal Process. **56**(6), 2334–2345 (2008)
131. Carin, L., Baraniuk, R., Cevher, V., Dunson, V., Jordan, M., Sapiro, G., Wakin, M.: Learning low-dimensional signal models. IEEE Signal Process. Mag. **28**(2), 39–51 (2011)
132. Blumensath, T.: Sampling and reconstructing signals from a union of linear subspaces. IEEE Trans. Inf. Theory **57**(7), 4660–4671 (2011)
133. Turaga, P., Veeraraghavan, A., Srivastava, A., Chellappa, R.: Statistical computations on Grassmann and Stiefel manifolds for image and video-based recognition. IEEE Trans. Pattern Anal. Mach. Intell. **33**(11), 2273–2286 (2011)
134. Hjort, N., Holmes, C., Müller, P., Walker, S. (eds.): Bayesian Nonparametrics. Cambridge University Press, Cambridge/New York (2010)
135. Cressie, N., Wikle, C.: Statistics for Spatio-Temporal Data. Wiley, Hoboken (2011)

136. Héas, P., Herzet, C., Mémin, E.: Bayesian inference of models and hyperparameters for robust optical-flow estimation. IEEE Trans. Image Process. **21**(4), 1437–1451 (2012)
137. MacKay, D.: Bayesian interpolation. Neural Comput. **4**(3), 415–447 (1992)
138. Tierney, L., Kadane, J.: Accurate approximations for posterior moments and marginal densities. J. Am. Math. Soc. **81**(393), 82–86 (1986)
139. Kass, R., Tierney, L., Kadane, J.: The validity of posterior expansions based on Laplace's method. In: Barnard, G.A., Geisser, S. (eds.) Bayesian and Likelihood Methods in Statistics and Econometrics, pp. 473–488. Elsevier Science, New York (1990)
140. Rue, H., Martino, S., Chopin, N.: Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. J. R. Stat. Soc. B **71**(2), 319–392 (2009)
141. Artana, G., Cammilleri, A., Carlier, J., Mémin, E.: Strong and weak constraint variational assimilations for reduced order fluid flow modeling. J. Comput. Phys. **231**(8), 3264–3288 (2012)
142. Talagrand, O., Courtier, P.: Variational assimilation of meteorological observations with the adjoint vorticity equation. I: theory. Q. J. R. Meteorol. Soc. **113**(478), 1311–1328 (1987)
143. Bain, A., Crisan, D.: Fundamentals of Stochastic Filtering. Springer, New York/London (2009)
144. Doucet, A., Godsil, S., Andrieu, C.: On sequential Monte Carlo sampling methods for Bayesian filtering. Stat. Comput. **10**, 197–208 (2000)

Non-linear Image Registration

Lars Ruthotto and Jan Modersitzki

Contents

1 Introduction 2006
2 The Mathematical Setting 2007
Variational Formulation of Image Registration 2008
Images and Transformation 2008
Length, Area, and Volume Under Transformation 2009
Distance Functionals 2010
Ill-Posedness and Regularization 2012
Elastic Regularization Functionals 2012
Hyperelastic Regularization Functionals 2014
Constraints 2016
Related Literature and Further Reading 2016
3 Existence Theory of Hyperelastic Registration 2018
Sketch of an Existence Proof 2019
Set of Admissible Transformations 2021
Existence Result for Unconstrained Image Registration 2023
4 Numerical Methods for Hyperelastic Image Registration 2032
Discretizing the Determinant of the Jacobian 2032
Galerkin Finite Element Discretization 2036
Multi-level Optimization Strategy 2039
5 Applications of Hyperelastic Image Registration 2042
Motion Correction of Cardiac PET 2042
Susceptibility Artefact Correction of Echo-Planar MRI 2044
6 Conclusion 2047
Cross-References 2048
References 2048

L. Ruthotto (✉)
Department of Mathematics and Computer Science, Emory University, Atlanta, Georgia, GA, USA
e-mail: lruthott@eos.ubc.ca

J. Modersitzki
Institute of Mathematics and Image Computing, University of Lübeck, Luebeck, Germany

© Springer Science+Business Media New York 2015

O. Scherzer (ed.), *Handbook of Mathematical Methods in Imaging*,
DOI 10.1007/978-1-4939-0790-8_39

Abstract

Image registration is to automatically establish geometrical correspondences between two images. It is an essential task in almost all areas involving imaging. This chapter reviews mathematical techniques for nonlinear image registration and presents a general, unified, and flexible approach. Taking into account that image registration is an ill-posed problem, the presented approach is based on a variational formulation and particular emphasis is given to regularization functionals motivated by mathematical elasticity. Starting out from one of the most commonly used linear elastic models, its limitations and extensions to nonlinear regularization functionals based on the theory of hyperelastic materials are considered. A detailed existence proof for hyperelastic image registration problems illustrates key concepts of polyconvex variational calculus. Numerical challenges in solving hyperelastic registration problems are discussed and a stable discretization that guarantees meaningful solutions is derived. Finally, two case studies highlight the potential of hyperelastic image registration for medical imaging applications.

1 Introduction

Image registration is an essential task in a variety of areas involving imaging techniques such as astronomy, geophysics, and medical imaging; see, e.g., [11, 27, 31, 45, 52, 55, 73] and references therein. The goal of image registration is to automatically find geometrical correspondences between two or more images acquired; see also Fig. 1 for a simplified example. Measurements may result from different times, from different devices, or perspectives. More specifically, one aims to find a *reasonable* transformation, such that a transformed version of the first image is *similar* to the second one. Mathematically, image registration is an ill-posed problem and is thus typically phrased as a variational problem involving data fit and regularization. The data fit or distance functional measures the similarity of the images. The regularization functional quantifies the reasonability of the transformation and can also be used to guarantee a mathematically sound formulation and to favor solutions that are realistic for the application in mind.

This chapter presents a comprehensive overview of mathematical techniques used for nonlinear image registration. A particular focus is on regularization techniques that ensure a mathematically sound formulation of the problem and allow stable and fast numerical solution.

Starting out from one of the most commonly used linear elastic models [10, 55], its limitations and extensions to nonlinear regularization functionals based on the theory of hyperelastic materials are discussed. A detailed overview of the available theoretical results is given and state-of-the-art numerical schemes as well as results for real-life medical imaging applications are presented.

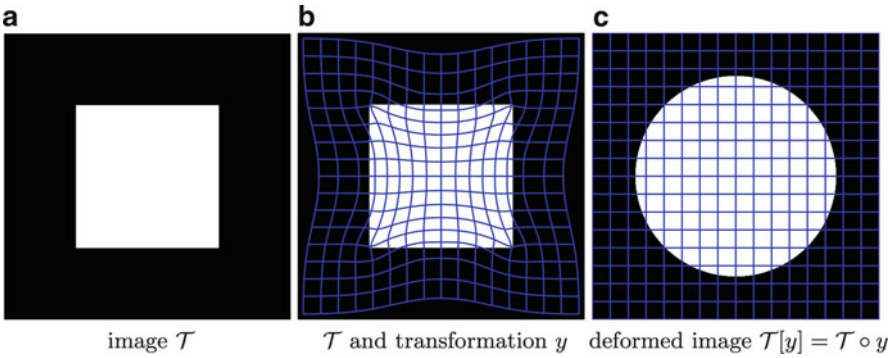


Fig. 1 Transforming a square into a disc: template image $\mathcal{T}$ is a *white square on black background* (*left*), reference image $\mathcal{R}$ is a *white disk on black background* (not displayed), the transformation y is visualized by showing a regular grid X as overlay on the deformed image $\mathcal{T}[y]$ (*right*) and as $y(X)$ on the template image (*center*); note that $\mathcal{T}[y](x) = \mathcal{T}(y(x))$

2 The Mathematical Setting

The purpose of this section is to provide a brief, conceptual introduction to a variational formulation of image registration, see section "Variational Formulation of Image Registration". For brevity, the presentation is restricted to a small subset of the various configurations of the variational problem and is limited to 3D problems which today is most relevant in medical imaging applications. Section "Images and Transformation" defines a continuous model for images and formalizes the concept of transformations y and transformed images $\mathcal{T}[y]$. Section "Length, Area, and Volume Under Transformation" discusses the transformation of length, area, and volume. These geometrical quantities play an important role for the computation of locally invertible transformations, which are important for many clinical applications.

After having discussed the general concepts on a rather broad level, specific building blocks of the variational framework are examined in more detail. Section "Distance Functionals" introduces two common choices of distance functionals $\mathcal{D}$, which measures the alignment of the images. After motivating the demand for regularization by illustrating the ill-posedness of the variational problem in section "Ill-Posedness and Regularization", an overview of elastic regularization functionals and a discussion of their limitations is given in section "Elastic Regularization Functionals".

Motivated by the limitations of elastic regularization functionals extensions to hyperelastic functionals, designed to ensure invertible and reversible transformations, are presented in section "Hyperelastic Regularization Functionals". Subsequently, section "Constraints" presents two examples that demonstrate how constraints $\mathcal{C}$ can be used to restrict set of admissible transformations $\mathcal{A}$, to favor

practically relevant solutions, and most importantly, to exclude practically infeasible solutions.

Finally, a comprehensive overview of related literature and suggestions for further readings are provided in section "Related Literature and Further Reading".

Variational Formulation of Image Registration

Given the so-called template image $\mathcal{T}$ and the so-called reference image $\mathcal{R}$, the image registration problem can be phrased as a variational problem. Following [55], an energy functional $\mathcal{J}$ is defined as a sum of the data fit $\mathcal{D}$ depending on the input images and the transformation and a regularization $\mathcal{S}$,

$$\mathcal{J}[y] := \mathcal{D}[\mathcal{T}, \mathcal{R}; y] + \mathcal{S}[y]. \tag{1}$$

The objective is then to find a minimizer y of $\mathcal{J}$ in a feasible set $\mathcal{A}$ to be described,

$$\min \mathcal{J}[y] \quad \text{for} \quad y \in \mathcal{A}. \tag{2}$$

Advantages of this formulation are its great modelling potential and its modular setting that allows to obtain realistic models tailored to essentially any particular application.

Images and Transformation

This section introduces a continuous image model, transformations, and discusses geometrical transformations of images. Although the image dimension is conceptually arbitrary, the description is restricted to 3D images for ease of presentation.

Definition 1 (Image). Let $\Omega \subset \mathbb{R}^3$ be a domain, i.e., bounded, connected, and open. An *image* is a one-time continuously differentiable function $\mathcal{T} : \mathbb{R}^3 \to \mathbb{R}$ compactly supported in Ω. The set of all images is denoted by $\mathrm{Img}(\Omega)$.

Note that already a simple transformation of an image such as a rotation requires a continuous image model. Therefore, a continuous model presents no limitation, even if only discrete data is given in almost all applications. Interpolation techniques are used to lift the discrete data to the continuous space; see, e.g., [55] for an overview. As registration is tackled as an optimization problem, continuously differentiable models can provide numerical advantages.

Definition 2 (Transformation). Let $\Omega \subset \mathbb{R}^3$ be a domain. A *transformation* is a function $y : \Omega \to \mathbb{R}^3$. For a differentiable transformations, the *Jacobian matrix* is denoted by

$$\nabla y := \left(\partial_1 y \mid \partial_2 y \mid \partial_3 y\right) := \begin{pmatrix} \partial_1 y_1 & \partial_2 y_1 & \partial_3 y_1 \\ \partial_1 y_2 & \partial_2 y_2 & \partial_3 y_2 \\ \partial_1 y_3 & \partial_2 y_3 & \partial_3 y_3 \end{pmatrix}.$$

The function $u : \Omega \to \mathbb{R}^3$ satisfying $y(x) = x + u(x)$ is called *displacement*, $\nabla y = \mathrm{I} + \nabla y$, where I denotes a diagonal matrix of appropriate size with all ones on its diagonal.

Example 1. An important class of transformations are *rigid transformations*. Let $b \in \mathbb{R}^3$ describes a translation and $Q \in \mathbb{R}^{3\times 3}$ a rotation matrix,

$$\begin{aligned} \mathcal{A}_{\text{rigid}}(\Sigma) := \{y : y(x) = Qx + b \text{ a.e. on } \Sigma, \\ b \in \mathbb{R}^3,\ Q \in \mathbb{R}^{3\times 3},\ Q^\top Q = I,\ \det Q = 1\}. \end{aligned}$$

A transformation y is rigid on Σ, if $y \in \mathcal{A}_{\text{rigid}}(\Sigma)$.

Given an image $\mathcal{T}$ and an invertible transformation y, there are basically the Eulerian or Lagrangian way to define the transformed version $\mathcal{T}[y]$; see [38, 55] for more details. In the Eulerian approach the transformed image is defined by

$$\mathcal{T}[y] : \mathbb{R}^3 \to \mathbb{R}, \quad \mathcal{T}[y](x) := \mathcal{T}(y(x)). \tag{3}$$

In other words, assigning position x in $\mathcal{T}[y]$ the intensity of $\mathcal{T}$ at position $y(x)$; see illustration in Fig. 1. The Lagrangian approach transports the information $(x, \mathcal{T}(x))$ to $(y(x), \mathcal{T}(x))$, where the first entry denotes location and the second intensity. In image registration, typically, the Eulerian framework is used. However, the Lagrangian framework can provide advantages in a constrained setting; see [38] for a local rigidity constrained and the discussion in section "Constraints".

Length, Area, and Volume Under Transformation

This section relates changes of length, area, and volume in the deformed coordinate system to the gradient of the transformation y. To simplify the presentation, it is assumed that y is invertible and at least twice continuously differentiable.

Using Taylor's formula [18], a local approximation of the transformation at an arbitrary point x in direction $v \in \mathbb{R}^3$ is given by

$$y(x + v) = y(x) + \nabla y(x)\, v + \mathcal{O}(|v|^2). \tag{4}$$

Choosing $v = he_i$, it can be seen that the columns of the gradient matrix $\nabla y(x)$ approximate the edges of the transformed reference element; see also Fig. 2. Thus, changes of length, area, and volume of a line segment L connecting x and $x + he_i$,

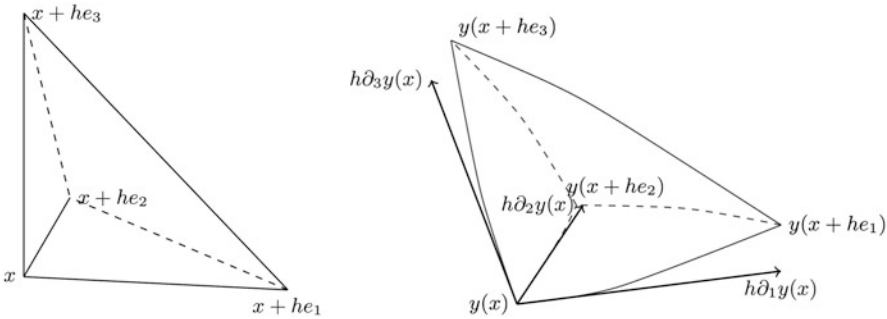


Fig. 2 Visualization of a deformation of a tetrahedron and the Taylor approximation of the transformation y. At an arbitrary point x a tetrahedron is spanned by the scaled Euclidean unit vectors (*left*). The edges of the deformed tetrahedron are approximated by the columns of the gradient $\nabla y(x)$ (*right*). Due to this approximation property, geometric quantities such as volume, area, and length can be related to the gradient matrix, its cofactor, and its determinant

a triangle T spanned by x, $x + he_j$, and $x + he_k$, and a tetrahedron P spanned by x and the three scaled Euclidean vectors are approximately

$$\begin{aligned}
\text{length}(y(L)) &= |y(x + he_i) - y(x)| = |\partial_i y| + \mathcal{O}(h^2), \\
\text{area}(y(T)) &= \frac{h^2}{2} \left|\partial_j y \times \partial_k y\right| + \mathcal{O}(h^2), \\
\text{vol}(y(P)) &= h^3 \ \det\Big(\partial_1 y \ \Big|\ \partial_2 y \ \Big|\ \partial_3 y\Big) + \mathcal{O}(h^2) = h^3 \det \nabla y + \mathcal{O}(h^2),
\end{aligned}$$

where $\times$ denotes the outer product in $\mathbb{R}^3$.

The cofactor matrix $\text{cof}\nabla y$ summarizes, column wise, the area changes of the three triangles spanned by x with two scaled Euclidean vectors each,

$$\text{cof}\nabla y = \Big(\partial_2 y \times \partial_3 y \ \Big|\ \partial_3 y \times \partial_1 y \ \Big|\ \partial_1 y \times \partial_2 y\Big) \in \mathbb{R}^{3\times 3}.$$

At this point, the importance of limiting the range of $\det \nabla y$ for many clinical applications should be stressed. Due to the inverse function theorem [25, Sect. 5], a transformation $y \in \mathcal{C}^1(\Omega, \mathbb{R}^3)$ is locally one-to-one if $\det \nabla y \neq 0$. From the above considerations, it can be seen that $\det \nabla y = 0$ indicates annihilation of volume and $\det \nabla y < 0$ indicates a change of orientation. Thus, controlling the range of $\det \nabla y$, or analogously the range of compression and expansion of volume, ensures local invertibility and thus is of utmost importance in practical applications.

Distance Functionals

The variational formulation of the image registration problem (2) characterizes an optimal transformation as a minimizer of the sum of a distance and a regularization

functional subject to optional constraints. These three building blocks will be introduced and discussed in the following sections.

Distance functionals mimic the observer's perception to quantify the similarity of images. In the following, two examples of distance measures will be presented and the reader is referred to [55] for further options. Emphasis is on a mass-preserving distance measure, which is an attractive option for the registration of density images such as Positron Emission Tomography (PET) images; see also section "Motion Correction of Cardiac PET".

A robust and widely use distance measure is the L^2-norm of the intensity differences commonly referred to as the *sum-of-squared-difference* (SSD),

$$\mathcal{D}^{\mathrm{SSD}}[y] = \mathcal{D}^{\mathrm{SSD}}[\mathcal{T}, \mathcal{R}; y] = \frac{1}{2}\int_{\Omega} (\mathcal{T}(y(x)) - \mathcal{R}(x))^2 dx. \tag{5}$$

For this distance measure to be effective, intensity values at corresponding points in the reference and template image are assumed to be equal. This assumption is often satisfied when comparing images of the same modality, commonly referred to as *mono-modal* registration. Hence, $\mathcal{D}^{\mathrm{SSD}}$ is a prototype of *mono-modal* distance functionals.

In some applications intensities at corresponding points differ conceptually. Typical examples of multimodal image registration include the fusion of different modalities like anatomy (such as CT or MRI) and functionality (such as SPECT or PET); see, e.g., [55] for further options and discussions.

Mass densities are another example for conceptional different image intensities. The necessity of mass-preserving transformations was first discussed in [29]. Recently, mass-preserving distance functionals were used to register images from Positron Emission Tomography (PET); see section "Motion Correction of Cardiac PET" and [14, 30, 60]. Another application is the correction of image distortions arising from field inhomogeneities in Echo-Planar Magnetic Resonance Imaging; see section "Susceptibility Artefact Correction of Echo-Planar MRI" and [15, 62].

Due to mass-preservation, change of volume causes change of intensities. Note that the simple model for transformed images (3) does not alter image intensities. Let $V \subset \Omega$ denotes a small reference volume. Ideally, the mass of $\mathcal{R}$ contained in V has to be equal to the mass of $\mathcal{T}$ contained in $y(V)$. Formally,

$$\int_V \mathcal{R}(x)dx = \int_{y(V)} \mathcal{T}(x)dx = \int_V \mathcal{T}(y(x)) \cdot \det \nabla y(x)\, dx, \tag{6}$$

where the second equality holds by the transformation theorem [18, p. 31f], assuming that the transformation is injective, continuously differentiable, orientation preserving, and that its inverse is continuous. A natural *mass-preserving* extension of the SSD distance functional thus reads

$$\mathcal{D}^{\mathrm{MP}}[y] := \frac{1}{2}\int_{\Omega} (\mathcal{T}(y(x)) \cdot \det \nabla y(x) - \mathcal{R}(x))^2\, dx. \tag{7}$$

Ill-Posedness and Regularization

A naive approach to image registration is to simply minimize the distance functional $\mathcal{D}$ with respect to y. However, this is an *ill-posed* problem [23,27,46,55]. According to Hadamard [41], a problem is *well-posed*, if there exists a solution, the solution is unique, and the solution depends continuously on the data. If one of these criteria is not fulfilled, the problem is called *ill-posed*.

In general, existence of a minimizer of the distance functional cannot be expected. Even the rather simple SSD distance functional (5) depends in a non-convex way on the transformation; for a discussion of more general distance functionals, see [55, p. 112]. If the space of plausible transformations is infinite dimensional, existence of solutions is thus critical.

A commonly used approach to bypass these difficulties and to ensure existence is to add a regularization functional that depends convexly on derivatives of the transformation [10, 67, 68]. This strategy is effective for distance functionals are independent of or depend convexly on the gradient of the transformation. This is the case in most applications of image registration, for instance, for the SSD in (5). However, further arguments are required for the mass-preserving distance functional in (7) due to the non-convex of the determinant; see Sect. 1.

It is also important to note that distance functionals may be affected considerably by noise; see also [55]. This problem is often (partially) alleviated by using regularization functionals based on derivatives of the transformation. This introduces a coupling between adjacent points, which can stabilize the problem against such local phenomena.

Note, even though regularization ensures existence of solutions, the objective functional depends in a non-convex way on the transformation and thus a solution is generally not unique as the following example illustrates.

Example 2. Consider an image of a plain white disc on a plain black background as a reference image and a template image showing the same disc after a translation. After shifting the template image to fit the reference image, both images are identical regardless of rotations around the center of the disc. Hence, there are infinitely many rigid transformations yielding images with optimal similarity. Regularization can be used to differentiate the various global optimizers.

Elastic Regularization Functionals

The major task of the regularization functional $\mathcal{S}$ is to ensure existence of solutions to the variational problem (2) and to ensure that these solutions are plausible for the application in mind. Therefore, regularization functionals that relate to a physical model are commonly used. In this section it is assumed that the transformation y is at least one time continuously differentiable and sufficiently measurable. A formal

definition of an elastic regularization functional and its associated function spaces is postponed to section "Hyperelastic Regularization Functionals".

The idea of elastic regularization can be traced back to the work of Fischler and Elschlager [28] and particularly the landmarking thesis of Broit [10]. The assumption is that the objects being imaged deform elastically. By applying an external force, one object is deformed in order to minimize the distance to the second object. This *external* force is counteracted by an *internal* force given by an elastic model. The registration process is stopped in an equilibrium state, that is, when external and internal forces balance.

In linear elastic registration schemes, the internal force is based on the displacement u and the *Cauchy strain tensor*

$$V = V(u) = (\nabla u + \nabla u^\top)/2,$$

which depends linearly on u. The *linear elastic* regularization functional is then defined as

$$\mathcal{S}^{\text{elas}}[u] = \int_\Omega \nu(\operatorname{trace} V)^2 + \mu \operatorname{trace}(V^2)\, dx,$$

where ν and μ are the so-called Lamé constants [18, 53].

Benefits and drawbacks of the model $\mathcal{S}^{\text{elas}}$ relate to the linearity of V in ∇u. Its simplicity has been exploited for the design of computationally efficient numerical schemes such as in [37]. A drawback is the limitation to small strains, that is, transformations with $\|\nabla u\| \ll 1$; see [10, 18, 55]. For large geometrical differences, the internal forces are modelled inadequately and thus solutions may become meaningless.

While the motivation in [10] is to stabilize registration against low image quality, elastic regularization also ensures existence of solutions in combination with many commonly used distance functionals. Another desired feature of this regularizer is that smooth transformations are favored, which is desirable in many applications.

In order to overcome the limitation to small strains, Yanovsky et al. [72] proposed an extension to nonlinear elasticity. They used the *Green-St.-Venant strain tensor*

$$E = E(u) = (\nabla u + \nabla u^\top + \nabla u^\top \nabla u)/2, \tag{8}$$

which is a quadratic in ∇u. Revisiting the computations of changes in length in section "Length, Area, and Volume Under Transformation", it can be shown that E penalized changes of lengths. The *quadratic elastic* regularization functional of [72] reads

$$\mathcal{S}^{\text{quad}}[u] = \int_\Omega \nu(\operatorname{trace} E)^2 + \mu \operatorname{trace}(E^2)\, dx. \tag{9}$$

While relaxing the small strain assumption, the functional is not convex as the following example shows. Thus, additional arguments are required to prove existence of a solution to the variational problem based on $\mathcal{S}^{\text{quad}}$.

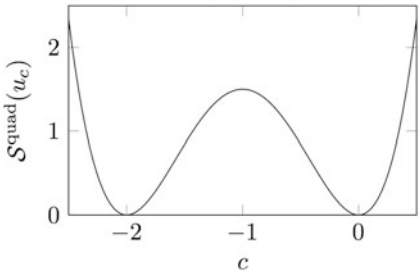


Fig. 3 Non-convexity of $\mathcal{S}^{\text{quad}}$ in (9) also used in Yanovsky et al. [72]. The value of $\mathcal{S}^{\text{quad}}$ is plotted for transformations $y_c := x + \text{diag}(c, c, 0)x$, $c \in [-2.5, 0.5]$ assuming parameters $\nu = \mu = 1$

Example 3 (Non-convexity of $\mathcal{S}^{\text{quad}}$). For $c \in \mathbb{R}$ a transformation y_c is defined as $y_c(x) := x + \text{diag}(c, c, 0)x$ such that $u_c(x) = \text{diag}(c, c, 0)x$. Thus, $E(u_c) = \text{diag}(c + c^2/2, c + c^2/2, 0)$. Assuming $\nu = \mu = 1$ in (9), it holds that $\mathcal{S}^{\text{quad}}[u_c] = 6(c + c^2/2)^2$.

Picking $c = 0$ and $c = -2$, it holds $\mathcal{S}^{\text{quad}}(u_{-2}) = \mathcal{S}^{\text{quad}}(u_0) = 0$; see also Fig. 3 for the landscape of $\mathcal{S}^{\text{quad}}(u_c)$. The non-convexity of $\mathcal{S}^{\text{quad}}$ with respect to y_c can then be seen by considering convex combinations of y_{-2} and y_0. Since $\mathcal{S}^{\text{quad}}(u_{-1}) = 3/2 > 0$, $\mathcal{S}^{\text{quad}}$ is not convex in y. Also note that y_{-1} is non-invertible and still has finite energy.

Another serious drawback of both the energies $\mathcal{S}^{\text{elas}}$ and $\mathcal{S}^{\text{quad}}$ is that both yield finite energy for transformations that are not one-to-one; see Example 4. In addition, even for invertible transformations, the above regularization functionals do not explicitly control tissue compression and expansion; see Example 5. Thus, unlimited compression is performed, if sufficient reduction of the distance functional can be gained.

Example 4. Consider the non-invertible trivial map $y(x) \equiv 0$, with $u(x) = -x$ and $\nabla u = -\text{I}$. Clearly, $V = 4E = -2\text{I}$ and $\mathcal{S}^{\text{elas}}[u] = 4\mathcal{S}^{\text{quad}}[u] = (9\nu + 3\mu)\text{vol}(\Omega)$. These regularizers give thus finite values that depend only on the Lamé constants and the volume of the domain Ω.

Example 5. For the sequence of transformations $y^{(n)}(x) := 2^{-n}x$, it follows $u(x) = (2^{-n} - 1)x$ and $\det \nabla y = 2^{-3n}$. For large n, this transformation heavily compresses volume. However, the components of ∇u are bounded below by -1, showing that arbitrary compressions are not detected properly.

Hyperelastic Regularization Functionals

Motivated by these observations, a desirable property of regularization functionals is rapid growth to infinity for $\det \nabla y \to 0^+$. In the Example 5, the Jacobian determinant is $\det \nabla y = 2^{-3n}$ indicating that the volume of a reference domain goes rapidly to zero with growing n. Controlling compression and expansion to

physically meaningful ranges has been a central theme in image registration; see, e.g., [34–36, 40, 58, 59].

Models for the so-called Ogden materials [4, 18] use an energy function that satisfies this important requirement,

$$\mathcal{S}^{\text{Ogden}}[y] = \int_\Omega \frac{1}{2}\alpha_\ell |\nabla u|^2 + \alpha_\text{a} \|\text{cof}\nabla y\|_F^2 + \alpha_\text{v}\, \psi_\text{O}(\det \nabla y) dx, \tag{10}$$

with regularization a parameters $\alpha_\ell, \alpha_\text{a}, \alpha_\text{v} > 0$, the *Frobenius norm* $\|A\|_F^2 = \sum_{i,j=1}^3 a_{ij}^2$, and the convex penalty $\psi_\text{O}(v) = v^2 - \log v$. Note that this penalty ensures $\lim_{v\to 0+} \psi_\text{O}(v) = \lim_{v\to\infty} \psi_\text{O}(v) = \infty$ and thus gives infinite energy for non-invertible transformations. For small strains, linear elasticity is an approximation to this regularization functional; see [18]. For image registration, this energy was introduced first by Droske and Rumpf [23].

While ensuring invertible transformations, the hyperelastic regularization functional (10) has drawbacks in the context of image registration: On the one hand, setting $\alpha_\text{a} > 0$ puts a bias towards transformations that reduce surface areas. On the other hand, setting $\alpha_\text{a} = 0$ prohibits the standard existence theory as presented in the next section. In addition, ψ_O is not symmetric with respect to inversion, that is, $\psi_\text{O}(v) \neq \psi_\text{O}(1/v)$. Since $\det \nabla y^{-1} = (\det \nabla y)^{-1}$ the value of the volume regularization term changes, when interchanging the role of template and reference image; see [17] for a discussion of the *inverse consistency* property.

To overcome these two drawbacks, the following *hyperelastic regularizer* as a slight modification of the Ogden regularizer (10) was introduced in [14],

$$\mathcal{S}^{\text{hyper}}[y] = \int_\Omega \frac{1}{2}\alpha_\ell(x) |\nabla u|^2 + \alpha_\text{a}(x)\phi_\text{c}(\text{cof}\nabla y) + \alpha_\text{v}(x)\, \psi(\det \nabla y) dx.$$

Making the regularization parameter spatially dependent is a small conceptual contribution with potentially big impact for many applications. The main idea is to replace ψ_O by a convex penalty ψ, fulfilling the growth condition, and satisfying $\psi(v) = \psi(1/v)$. More precisely, the convex functions

$$\phi_\text{c}(C) = \sum_{i=1}^3 \max\left\{\sum_{j=1}^3 C_{ji}^2 - 1, 0\right\}^2 \text{ and } \psi(v) = \tfrac{(v-1)^4}{v^2}$$

were suggested in [14]. The penalty ϕ_c is a convexification of the distance to orthogonal matrices and the penalty $\psi(v)$ is essentially a polynomial of degree two, thus guaranteeing $\det \nabla y \in L^2$, which is important in mass-preserving registration. If $\det \nabla y \in L^1$ suffices, ψ can be replaced by its square root. Note that ϕ_c is designed to avoid the preference for area shrinkage in (10). A convexification is needed for the existence theory, which makes ϕ_c blind against shrinkage, allowing only penalization of area expansion.

Constraints

It is often beneficial to include additional prior knowledge or expectations on the wanted transformation into the model. Two case studies will show exemplarily how to restrict the set of admissible transformations. The first one is related to volume constraints [35, 36, 59] and the second one to local rigidity constraints [38, 65].

Volume-preserving constraints are particularly important in a setting such as monitoring tumor growth, where a change of volume due to the mathematical model can be critical. In addition, registration problems requiring highly nonlinear deformations, e.g., arising in abdominal or heart imaging, can be stabilized by restricting volume change to a certain class.

Constraining compression and expansion of volume has been a central topic in image registration over many years. Following [36], a formal description reads

$$\kappa_m(x) \leq \mathcal{C}^{\mathrm{VC}}[y](x) := \det \nabla y(x) \leq \kappa_M(x) \text{ for } x \in \Omega,$$

where κ_m and κ_M are given bounds. Volume preservation was enforced by using penalties, equality constraints such as $\kappa_m = \kappa_M \equiv 1$, or box constraints; see, e.g., [34–36, 40, 58, 59].

Another example for image-based constraints is *local rigidity*. Local rigidity constraints can improve the model in applications in which images show bones or other objects that behave approximately rigidly, such as head-neck images, motion of prostate in the pelvis cage, or the modelling of joints. Rigidity on a subdomain $\Sigma \subset \Omega$ can be enforced by setting

$$\mathcal{C}^{\mathrm{LR}}[y, \Sigma] := \mathrm{dist}(y, \mathcal{A}_{\mathrm{rigid}}(\Sigma)\}) = 0;$$

see also Example 1. This formulation can be extended to multiple regions; see [38, 65].

Based on the discussion in [38], the Lagrangian framework has the advantage that the set Σ does not depend on y and thus does not need to be tracked; see [54] for a tracked Eulerian variant. Tracking the constrained regions may add discontinuities to the registration problem. In the case of local rigidity constraints, constraint elimination is an efficient option and results linear constraints. However, the Lagrangian framework involves $\det \nabla y$ in the computations of $\mathcal{D}$ and $\mathcal{S}$; see [38] for details.

Related Literature and Further Reading

Due to its many applications and influences from different scientific disciplines, it is impossible to give a complete overview of the field of image registration. Therefore this overview is restricted to some of the works, which were formative for the field or are relevant in the scope of this chapter. For general introductions to

image registration, the reader is referred to the textbooks [33, 53, 55]. In addition, the development of the field has been documented in a series of review articles that appeared in the last two decades; see [11, 27, 31, 45, 52, 73].

A milestone for registration of images from different imaging techniques, or *modalities*, was the independent discovery of the *mutual information* distance functional by Viola and Wells [71] and Collignon et al. [19]. In this information theoretic approach, the reference and template images are discretized and interpreted as a sequence of grey values. The goal is to minimize the entropy of the joint distribution of both sequences. Due to its generality, mutual information distance measures can be used for a variety of image modalities. However, it is tailored to discrete sequences and thus the mutual information of two continuous images is not always well defined; see [42]. Furthermore, a local interpretation of the distance is impossible. To overcome these limitations, a *morphological* distance functional [23] and an approach based on *normalized gradient fields* [56] were derived for multimodal registration. Both approaches essentially use image gradients to align corresponding edges or level sets.

In addition to the already mentioned elastic and hyperelastic regularization functionals, another example is the *fluid registration* by Christensen [16]. The difference between fluids and elastic materials lies in the model for the inner force. In contrast to elastic materials, fluids have no memory about their reference state. Thus, the fluid registration scheme allows for large, nonlinear transformations while preserving continuity and smoothness. However, similar limitations arise from the fact that volume changes are not monitored directly.

It is well known that nonlinear registration schemes may fail if the geometrical difference between the images is too large. To overcome this, a preregistration with a rigid transformation model is usually employed. The above regularizers are, however, not invariant to rigid transformations. To overcome this limitation, the *curvature* regularization functional, based on second derivatives, was proposed [26,43].

Similar to the linearized elastic regularization functional, fluidal or curvature regularized schemes fail to detect non-invertible transformations. As an alternative to hyperelastic schemes, the invertibility of the solution is to restrict the search for a plausible transformation to the set of diffeomorphisms as has been originally suggested by Trouvé in 1995 [69] and resulted in important works [3, 5, 70]. A transformation y is diffeomorphic, if y and y^{-1} exist and are continuously differentiable. The existence of an inverse transformation implies that diffeomorphisms are one-to-one. While invertibility is *necessary* in most problems, it is not always *sufficient* as the Example 5 indicates: for large n the size of the transformed domain has no physical meaning; see [55] for further discussions.

Image registration is closely related to the problem of determining *optical flow*, which is a velocity field associated with the motion of brightness pattern in an image sequence. First numerical schemes have been proposed by Horn and Schunck [48] and Lucas and Kanade [51] in 1981. In their original formulations, both approaches assume that a moving particle does not change its intensity. This gives one scalar constraint for each point. Hence, as image registration, determining the flow field is an under-determined problem. The under-determinedness is addressed differently

by both approaches. The method by Horn and Schunck generates dense and globally smooth flow fields by adding a regularization energy. Lucas and Kanade, in contrast, suggest to smooth the image sequence and compute the flow locally without additional regularization. A detailed comparison and a combination of the benefits of both approaches is presented in [12].

There are many variants of optical flow techniques. For example, Brune replaced the brightness constancy by a mass-conservation constraint; see [13]. This gives the optical flow pendant to mass-preserving image registration schemes; see [30, 60, 61] and sections "Distance Functionals" and "Motion Correction of Cardiac PET". Further, rigidity constraints have been enforced in optical flow settings in [50].

One difference between image registration and optical flow lies in the available data. In optical flow problems, motion is typically computed based on an image sequence. Thus, the problem is space and time dependent. If the time resolution of the image sequence is sufficiently fine, geometrical differences between successive frames can be assumed to be small and displacements can be linearized. In contrast to that, in image registration the goal is typically to compute *one* transformation that relates two images with typically substantial geometrical differences.

The problem of mass-preserving image registration is closely related to the famous Monge–Kantorovich problem of *Optimal Mass Transport* (OMT) [1, 6, 24, 49]. It consists of determining a transport plan to transform one density distribution into another. To this end, a functional measuring transport cost is minimized over all transformations that satisfy $\mathcal{D}^{\mathrm{MP}}[y] = 0$ in (7). A key ingredient is the definition of the cost functional which defines *optimality*. Clearly there is an analogy the cost functional in OMT and the regularization functional in mass-preserving image registration.

3 Existence Theory of Hyperelastic Registration

The variational image registration problem (2) is *ill-posed* problem in the sense of Hadamard; see section "Ill-Posedness and Regularization" and [27, 55]. As Example 2 of a rotating disc suggests, uniqueness can in general not be expected. However, hyperelastic image registration has a sound existence theory which will be revisited in this section.

The main result of this section is the existence theorem for unconstrained hyperelastic image registration that includes all commonly used distance measures. It is further shown that solutions satisfy $\det \nabla y > 0$, which is necessary in many applications. The theory is complicated due to the non-convexity of the determinant mapping as has been pointed out by Ciarlet [18, p.138f]. Using his notation $F = \nabla u$: *"The lack of convexity of the stored energy function with respect to the variable F is the root of a major difficulty in the mathematical analysis of the associated minimization problem."*

This section organizes as follows: A sketch of the existence proof is given in section "Sketch of an Existence Proof". Formal definitions of involved function spaces, the set of admissible transformations, and the hyperelastic regularization

functional $\mathcal{S}^{\text{hyper}}$ are given in section "Set of Admissible Transformations". Finally, an existence result is shown in section "Existence Result for Unconstrained Image Registration".

Sketch of an Existence Proof

This section gives an overview of the existence theory of hyperelastic image registration problems. The goal is to outline the major steps, give intuition about their difficulties, and introduce the notation. The presentation is kept as informal as possible. The mentioned definitions, proofs, and theorems from variational calculus can be found, for instance, in [25, Ch. 8] or [18, 20].

The objective functional in (1) can also be written as an integral over a function $f : \Omega \times \mathbb{R}^3 \times \mathbb{R}^{3\times 3} \to \mathbb{R}$:

$$\mathcal{J}[y] = \int_\Omega f\big(x, y(x), \nabla y(x)\big)\, dx, \tag{11}$$

where the dependence on the template and reference images is hidden for brevity.

In the following it is assumed that f is continuously differentiable and measurable in x. This is granted for the part related to the hyperelastic regularization functional. For the part related to the distance functional, this can be achieved by using a continuous image model $\mathcal{T}$ and $\mathcal{R}$ and careful design of $\mathcal{D}$.

Typically, $\mathcal{J}$ is bounded below by a constant $m := \inf_y \mathcal{J}[y]$. The goal is to find a minimizer in a certain set of admissible transformations $\mathcal{A}$, which is a subset of a suitable Banach space X. Since X must be complete under *Lebesgue norms*, our generic choice is a *Sobolev space* $W^{1,p}(\Omega, \mathbb{R}^3)$, with $p \geq 1$, rather than $C^1(\Omega, \mathbb{R}^3)$.

Existence of minimizers is shown in three steps. Firstly, a *minimizing sequence* $\{y^k\}_{k\in\mathbb{N}} \subset \mathcal{A}$ is constructed, that is, a sequence with $\lim_{k\to\infty} \mathcal{J}(y^k) = m$ containing a convergent subsequence. Secondly, it is shown that its limit y^* is a minimizer of $\mathcal{J}$, in other words, $\mathcal{J}[y^*] = m$. Finally, it is verified that y^* actually belongs to the admissible set $\mathcal{A}$.

Key ingredients of the existence proof are *coercivity* and *lower semicontinuity*.

In a finite dimensional setting, the coercivity inequality

$$\mathcal{J}[y] \geq C\ \|y\|_X + K \text{ for all } y \in \mathcal{A} \text{ and constants } C > 0 \text{ and } K \in \mathbb{R}, \tag{12}$$

guaranties that $\{y^k\}_k$ lies in a bounded and closed set. From the lower semicontinuity it follows

$$y^k \to y^* \Rightarrow m = \lim_{k\to\infty} \mathcal{J}[y^k] \geq \mathcal{J}[y^*] \geq m,$$

and thus y^* is a minimizer. However, $W^{1,p}(\Omega, \mathbb{R}^3)$ is infinite dimensional. Hence, bounded and closed sets are in general not compact and the existence of a

norm-convergent subsequence cannot be assumed. Thus, less strict definitions of convergence have to be used.

Note that $W^{1,p}(\Omega, \mathbb{R}^3)$ is *reflexive* for $p > 1$. For these cases, it can be shown that there exists a subsequence that converges in the *weak topology*, i.e., with respect to all continuous, linear functionals on $W^{1,p}(\Omega, \mathbb{R}^3)$. Consequently, if $\mathcal{J}$ fulfills a coercivity condition (12), a bounded minimizing sequence can be constructed which yields weakly converging subsequence.

The second part, namely showing that the weakly convergent sequence actually converges to a minimizer, is in general more involved. Lower continuity of $\mathcal{J}$ with respect to weak convergence typically requires convexity arguments. In many examples of variational calculus, the integrand f of $\mathcal{J}$ is convex in ∇y. Thus, the integrand f can be bounded below by a linearization in the last argument around ∇y. Using the weak convergence of $\nabla y_k \rightharpoonup \nabla y^*$ and noting that $y_k \to y^*$ strongly in L^2 by a *compact embedding*, the linear term vanishes and lower semicontinuity follows; see Theorem 2.

When using hyperelastic regularization or mass-preserving distance functionals, $\mathcal{J}$ also depends on the Jacobian determinant $\det \nabla y$. Thus, the dependence of f on ∇y is non-convex and further arguments are required to obtain lower semicontinuity. To overcome this complication, one can follow the strategy suggested by Ball [4]. The idea is to introduce a splitting in $(y, \mathrm{cof}\nabla y, \det \nabla y)$ and show convergence of the sequence

$$\{(y^k, \mathrm{cof}\nabla y^k, \det \nabla y^k)\}_k \subset W^{1,p}(\Omega, \mathbb{R}^3) \times L^q(\Omega, \mathbb{R}^{3\times 3}) \times L^r(\Omega, \mathbb{R}) =: X$$

where the exponents $q > 0$ and $r > 0$ are appropriately chosen. A coercivity inequality in the $\|\cdot\|_X$-norm yields a weakly converging subsequence as outlined above. Extending f by additional arguments, it follows

$$\begin{aligned} g &: \Omega \times \mathbb{R}^3 \times \mathbb{R}^{3\times 3} \times \mathbb{R}^{3\times 3} \times [0, \infty) \to \mathbb{R} \\ &\text{with } g(x, y, \nabla y, \mathrm{cof}\nabla y, \det \nabla y) := f(x, y, \nabla y) \end{aligned}$$

is measurable in x and continuously differentiable. Most importantly, g is convex in the last three arguments. Consequently, weak lower semicontinuity of $\mathcal{J}$ in X can be shown, i.e., $(y^k, \mathrm{cof}\nabla y^k, \det \nabla y^k) \rightharpoonup (y^*, H^*, v^*)$ implies

$$\lim_{k\to\infty} \int_\Omega g(x, y^k, \nabla y^k, \mathrm{cof}\nabla y^k, \det \nabla y^k)\, dx \geq \int_\Omega g(x, y^*, \nabla y^*, H^*, v^*)\, dx = m.$$

To undo the splitting, the identifications $H^* = \mathrm{cof}\nabla y^*$ and $v^* = \det \nabla y^*$ must be shown. This is uncritical if y is sufficiently measurable, i.e., p is greater than the space dimension d. Note that p is also the degree of the polynomial $\det \nabla y$. For hyperelastic regularization, it is possible to show existence under even weaker assumptions, i.e., for $d = 3$ and $p = 2$. Key is the identity

$$\det \nabla y \cdot I = \nabla y^{\top} \mathrm{cof} \nabla y.$$

It will be shown that $H^* = \mathrm{cof}\nabla y^*$ and $v^* = \det \nabla y^*$ for exponents $q = 4$ and $r = 2$; see Theorem 4.

A final note is related to the positivity of the determinant. To assure that the Jacobian determinant of a solution is strictly positive almost everywhere, sufficient growth of $\mathcal{J}$ is required, i.e., $\mathcal{J}[y] \to \infty$ for $\det \nabla y \to 0^+$. Note, that $\det \nabla y^* > 0$ is not obvious. For example, $k^{-1} \to 0$ although $k^{-1} > 0$ for all k.

Set of Admissible Transformations

As pointed out, regularization functionals involving $\det \nabla y$ considerably complicate existence theory due to the non-convexity of the determinant mapping. Thus, a careful choice of function spaces and the set of admissible transformations is important for the existence theory to be investigated in the next section. The goal of this section is to define the function spaces and the set of admissible transformations underlying the hyperelastic regularization functional (10).

To begin with, the problems arising from the non-convexity of the determinant are addressed. One technique to bypass this issue is a splitting into the transformation, cofactor and determinant and study existence of minimizing sequences in a product space. Here the product space

$$X := W^{1,2}(\Omega, \mathbb{R}^3) \times L^4(\Omega, \mathbb{R}^{3\times 3}) \times L^2(\Omega, \mathbb{R}) \tag{13}$$

with the norm

$$\|(y, \mathrm{cof}\nabla y, \det \nabla y)\|_X := \|y\|_{W^{1,2}(\Omega,\mathbb{R}^3)} + \|\mathrm{cof}\nabla y\|_{L^4(\Omega,\mathbb{R}^{3\times 3})} + \|\det \nabla y\|_{L^2(\Omega,\mathbb{R})} \tag{14}$$

is considered. The space X is reflexive and the norms are motivated by the fact that the penalty functions ϕ_c and ψ presented in section "Hyperelastic Regularization Functionals" are essentially polynomials of degree four and two, respectively. The splitting is well defined for transformations in the set

$$\mathcal{A}_0 := \{y : (y, \mathrm{cof}\nabla y, \det \nabla y) \in X, \ \det \nabla y > 0 \text{ a.e.}\}. \tag{15}$$

To satisfy a coercivity inequality, boundedness of the overall transformation is required. In the standard theory of elasticity, this is typically achieved by imposing boundary conditions; see [4, 18]. In the image registration literature, a number of conditions, for instance, *Dirichlet*, *Neumann*, or *sliding* boundary conditions were used; see [53].

Since finding boundary conditions that are meaningful for the application in mind often is difficult, the existence proof in section "Existence Result for Unconstrained Image Registration" is based on the following arguments. Recall that images are

compactly supported in a domain Ω, which can be bounded by a constant $M \in \mathbb{R}$. Since images vanish outside Ω, reasonable displacements are also bounded by a constant only depending on Ω; see also discussions in [14, 58],

$$\|y\|_\infty \leq M + \text{diam}(\Omega).$$

For larger displacements, the distance functional is constant in y as the template image vanishes and thus no further reduction of the distance functional can be achieved. In this case, the regularization functional drives the optimization to a local minimum. Working with the supremum of y, however, complicates our analysis as the transformations in the non-reflexive space L^∞. Therefore, an averaged version of a boundedness condition can be used:

$$\tfrac{1}{\text{vol}(\Omega)} \left| \int_\Omega y(x)dx \right| \leq M + \text{diam}(\Omega).$$

This leads to the following characterization of admissible transformations and definition of the hyperelastic regularization energy.

Definition 3 (Hyperelastic Regularization Functional). Let $\Omega \subset \mathbb{R}^3$ be a domain bounded by a constant $M > 0$, $\alpha_i \in \mathcal{C}^1(\Omega, \mathbb{R}^+)$ be regularization parameters with $\alpha_i(x) \geq a_i > 0$ for all x and $i = l, a, v$, and $\mathcal{A}_0$ as in (15). The set of *admissible transformations* is

$$\mathcal{A} = \{y \in \mathcal{A}_0 \ : \ \left| \int_\Omega y(x)dx \right| \leq \text{vol}(\Omega)(M + \text{diam}(\Omega))\}.$$

Then the *hyperelastic regularization* functional is defined as

$$\mathcal{S}^{\text{hyper}} : \mathcal{A} \to \mathbb{R}^+, \qquad \mathcal{S}^{\text{hyper}}[y] = \mathcal{S}^{\text{length}}[y] + \mathcal{S}^{\text{area}}[y] + \mathcal{S}^{\text{vol}}[y], \tag{16}$$

with

$$\mathcal{S}^{\text{length}}[y] = \int_\Omega \alpha_\ell(x) \, |\nabla u(x)|^2 dx, \tag{17}$$

$$\mathcal{S}^{\text{area}}[y] = \int_\Omega \alpha_{\text{a}}(x) \, \phi_{\text{c}}(\text{cof}\nabla y(x))dx, \tag{18}$$

$$\mathcal{S}^{\text{vol}}[y] = \int_\Omega \alpha_{\text{v}}(x) \, \psi(\det \nabla y(x))dx, \tag{19}$$

and the convex functions $\phi_{\text{c}} : \mathbb{R}^{3\times 3} \to \mathbb{R}^+$ and $\psi : \mathbb{R} \to \mathbb{R}^+$,

$$\phi_{\text{c}}(C) = \sum_{i=1}^3 \max\left\{\sum_{j=1}^3 C_{ji}^2 - 1, 0\right\}^2 \text{ and } \psi(v) = \tfrac{(v-1)^4}{v^2}.$$

Existence Result for Unconstrained Image Registration

This section shows that hyperelastic regularization is sufficient to guarantee the existence of optimal transformations for unconstrained image registration problems and that the Jacobian determinants of minimizers of $\mathcal{J}$ in (2) are positive almost everywhere.

There are of numerous ways of showing existence. One option is to follow John Ball's proof for hyperelastic materials [4, 18, 20]. In [14], existence is obtained by verifying that the modified hyperelastic regularization functional still satisfies the assumptions of Ball's theorem which yields a very brief and concise result.

In this section, a new and different approach to the existence theory is given. Its main purpose is to give insight into the machinery behind Ball's proof and to prove existence in a simplified setting using only moderate theoretical tools. For the sake of clarity, some technical difficulties are avoided by making additional assumptions that are no limitation for most applications and also discussed below:

A1: The distance functional is of the form

$$\mathcal{D}[y] = \int_\Omega g_D(x, y(x), \mathrm{cof}\nabla y(x), \det \nabla y(x))\, dx,$$

where $g_D : \Omega \times \mathbb{R}^3 \times \mathbb{R}^{3\times 3} \times \mathbb{R} \to \mathbb{R}^+$ is continuously differentiable, convex in its last argument, and measurable in x.

A2: The regularization parameters satisfy $\alpha_\ell = \alpha_\mathrm{a} = \alpha_\mathrm{v} = 1$.

A3: With $\mathrm{Id}(x) = x$ for all x, it holds $\mathcal{J}(\mathrm{Id}) < \infty$.

A4: The boundary $\partial\Omega$ of the domain Ω is in $\mathcal{C}^1$.

(A1) holds for a large class of distance functionals, including the ones in [55], the mass-preserving distance functional (7), and distance functionals in the Lagrangian frame. Differentiability of g_D requires a differentiable image model and is assumed for simplicity. It can be weakened to the concept of *Carathéodory* functions; see [14, 20]. (A2) is made for ease of presentation. In general, there exists a constant $a > 0$ such that for all three penalty functions $a_i(x) \geq a$. Hence, (A2) present no loss of generality. For practical applications, (A3) is satisfied as the distance functional is finite for the initial images. (A4) allows to use a generalized Poincaré inequality [25, Sect. 5.8] and an integration by parts formula [25, Sect. C.1].

The main goal of this section is to prove the following theorem.

Theorem 1 (Existence of Optimal Transformations). *Given are images $\mathcal{T}, \mathcal{R} \in \mathrm{Img}(\Omega)$, a distance functional $\mathcal{D}$, $\mathcal{S}^{\mathrm{hyper}}$ and $\mathcal{A}$ as in Definition 3. Let the assumptions (A1)–(A4) be satisfied. There exists at least one minimizer $y^* \in \mathcal{A}$ of $\mathcal{J} = \mathcal{D} + \mathcal{S}^{\mathrm{hyper}}$.*

The proof is divided into a number of Lemmata following the sketch presented in section "Sketch of an Existence Proof". To begin with, coercivity of the objective functional in the space X as in (13) is shown. This means that $\mathcal{J}$ grows sufficiently

fast with respect to the $\|\cdot\|_X$ norm; see (14). This will be essential to bound the norm of the minimizing sequence.

Lemma 1 (Coercivity of $\mathcal{J}$). *Under the assumptions of Theorem 1, the functional $\mathcal{J}$ satisfies a coercivity inequality:* $\exists\, C > 0,\ K \in \mathbb{R}\ \forall\ y \in \mathcal{A}$:

$$\mathcal{J}[y] \geq K + C(\|y\|^2_{W^{1,2}} + \|\mathrm{cof}\nabla y\|^4_{L^4} + \|\det \nabla y\|^2_{L^2}).$$

Proof. Let $y \in \mathcal{A}$ be an admissible transformation. Since $\mathcal{D} \geq 0$ by assumption, coercivity of $\mathcal{J}$ follows from the coercivity of the regularization term

$$\mathcal{J}[y] \geq \mathcal{S}^{\text{hyper}}[y] = \int_\Omega \|\nabla y - \mathrm{I}\|^2_{\text{Fro}} + \phi_{\text{c}}(\mathrm{cof}\nabla y) + \psi(\det \nabla y)\, dx.$$

The penalties ϕ_{c} and ψ are designed such that there exist constants $C_1 > 0$ and $K_1 \in \mathbb{R}$ such that

$$\mathcal{J}[y] \geq K_1 + C_1 \left(\int_\Omega \|\nabla y - \mathrm{I}\|^2_{\text{Fro}} + \|\mathrm{cof}\nabla y\|^4_{\text{Fro}} + (\det \nabla y)^2\right)\, dx.$$

Using that $(a-b)^2 \geq \frac{1}{2}a^2 - b^2$ for all $a, b \in \mathbb{R}$, it follows that

$$\mathcal{J}[y] \geq K_1 + C_1 \left(\int_\Omega \tfrac{1}{2}\|\nabla y\|^2_{\text{Fro}} - \|\mathrm{I}\|^2_{\text{Fro}} + \|\mathrm{cof}\nabla y\|^4_{\text{Fro}} + (\det \nabla y)^2\right)\, dx.$$

With constants $C_2 > 0$ and $K_2 \in \mathbb{R}$, one obtains the inequality

$$\mathcal{J}[y] \geq K_2 + C_2 \left(\|\nabla y\|^2_{L^2} + \|\mathrm{cof}\nabla y\|^4_{L^4} + \|\det \nabla y\|^2_{L^2}\right). \tag{20}$$

It remains to be shown that the norm of ∇y in L^2 bounds the norm of the transformation y. To this end, assumption (A4) is used and a generalized Poincaré inequality [25, Sect. 5.8] is applied. In addition, the boundedness of the mean of transformations in $\mathcal{A}$ is used:

$$\begin{aligned}\|\nabla y\|^2_{L^2} &\geq C_3 \left\| y - \tfrac{1}{|\Omega|}\int_\Omega y(x)dx \right\|^2_{L^2} = C_3\|y\|^2_{L^2} - |\Omega| \left(\tfrac{1}{|\Omega|}\left|\int_\Omega y(x)dx\right|\right)^p \\ &\geq C_3\|y\|^2_{L^2} - |\Omega| \left(\tfrac{1}{|\Omega|}\int_\Omega |y(x)|dx\right)^2.\end{aligned}$$

Due to the assumption on $\mathcal{A}$, there exists a constant $K_3 \in \mathbb{R}$ independent on y such that

$$\|\nabla y\|^2_{L^2} \geq K_3 + C_3\|y\|^2_{L^2}. \tag{21}$$

Combining (20) and (21) and introducing cosmetics constants $C > 0$ and $K \in \mathbb{R}$ one obtains the desired growth condition

$$\mathcal{J}[y] \geq K + C(\|y\|^2_{W^{1,2}} + \|\mathrm{cof}\nabla y\|^4_{L^4} + \|\det \nabla y\|^2_{L^2}).$$

Since y was chosen arbitrarily, the assertion follows.

The following theorem establishes lower semicontinuity of the objective functional $\mathcal{J}$. It is a modified version of Theorem 1 of Evans [25, Sec. 8.2]. In contrast to the version of Evans, the next theorem does not necessarily assume that ζ is replaced by ∇y. However, it is assumed that $f \geq 0$, $y^k \to y^*$ strongly, and $\zeta^k \to \zeta^*$ weakly.

Theorem 2 (A General Lower Semicontinuity Result). *Let $\Omega \subset \mathbb{R}^3$ be a domain. Let $f : \Omega \times \mathbb{R}^3 \times \mathbb{R}^N \to [0, \infty]$ be continuously differentiable and $f(\cdot, y, \zeta)$ be measurable for every fixed y and ζ. Let $f(x, y, \cdot)$ be convex. For two sequences $y^k \to y^*$ in $L^p(\Omega, \mathbb{R}^3)$ with $p \geq 1$ and $\zeta^k \rightharpoonup \zeta^*$ in $L^q(\Omega, \mathbb{R}^N)$ with $q \geq 1$, it holds*

$$\liminf_{k\to\infty} \int_\Omega f(x, y^k(x), \zeta^k(x))dx \geq \int_\Omega f(x, y^*(x), \zeta^*(x))dx.$$

Proof. The proof basically follows the one of Theorem 1 [25, Sec. 8.2]. Simplifications arise from the fact that the sequences $\{y^k\}_k$ and $\{\zeta^k\}_k$ converge by assumption and $f \geq 0$. Another difference is that the second argument of f is used for a general sequence ζ^k and its limit ζ^* instead of ∇y^k and ∇y^*, respectively. The proof is organized in three steps.

Step 1: Since $0 \leq f$, it follows

$$0 \leq m := \liminf_{k\to\infty} \int_\Omega f(x, y^k(x), \zeta^k(x))dx,$$

thus there exists a (sub)-sequence $\{y^k, \zeta^k\}_k$ with

$$0 \leq m = \lim_{k\to\infty} \int_\Omega f(x, y^k(x), \zeta^k(x))dx.$$

Step 2: The strong convergence $y^k \to y^*$ in L^2 implies convergence almost everywhere. Thus, Egoroff's Theorem [25, Sect. E.2] can be used, which implies that for each $\epsilon > 0$ there exists a set $\Omega^1_\epsilon \subset \Omega$ with $\mathrm{vol}(\Omega \setminus \Omega^1_\epsilon) \leq \epsilon$ such that

$$y^k \to y^* \text{ uniformly on } \Omega^1_\epsilon. \tag{22}$$

Next, Ω^1_ϵ is reduced by regions where y^* or ζ^* are too large. Consider $\Omega_\epsilon := \Omega^1_\epsilon \cap \{x \in \Omega : |y^*(x)| + |\zeta^*(x)| \leq \epsilon^{-1}\}$. It holds that

$$\mathrm{vol}(\Omega \setminus \Omega_\epsilon) \to 0 \text{ for } \epsilon \to 0. \tag{23}$$

Step 3: From $f \geq 0$ and the convexity of f in the last argument (see also [25, Sect. B.1]), it follows for each fixed k that

$$\int_{\Omega} f(x, y^k, \zeta^k)dx \geq \int_{\Omega_\epsilon} f(x, y^k, \zeta^k)dx \\ \geq \int_{\Omega_\epsilon} f(x, y^k, \zeta^*)dx + \int_{\Omega_\epsilon} d_\zeta f(x, y^k, \zeta^*)\,(\zeta^k - \zeta^*)dx. \tag{24}$$

Due to the uniform convergence of $\{y^k\}_k$, see (22) and (23), it follows

$$\lim_{k\to\infty} \int_{\Omega_\epsilon} f(x, y^k, \zeta^k) = \int_{\Omega_\epsilon} f(x, y^*, \zeta^k). \tag{25}$$

Since also $d_\zeta f(x, y_k, \zeta^*) \to d_\zeta f(x, y^*, \zeta^*)$ uniformly on Ω_ϵ and $\zeta^k \rightharpoonup \zeta^*$ it holds that

$$\lim_{k\to\infty} \int_{\Omega_\epsilon} d_\zeta f(x, y^k, \zeta^*)(\zeta^k - \zeta^*)dx = 0. \tag{26}$$

Combining (25), (26), and (24), it follows

$$m = \lim_{k\to\infty} \int_{\Omega} f(x, y^k, \zeta^k)dx \geq \int_{\Omega_\epsilon} f(x, y^*, \zeta^*)dx.$$

This inequality holds for all $\epsilon > 0$. Using $f \geq 0$, the measurability of f, the Monotone Convergence Theorem [25, Sect. E.3] and letting $\epsilon \to 0$, it follows that

$$m \geq \int_{\Omega} f(x, y^*, \zeta^*)dx,$$

which concludes the proof.

Having shown the above general lower semicontinuity result, it is now verified that the objective functional $\mathcal{J}$ satisfies the assumptions that have been made.

Lemma 2 (Weak Lower Semicontinuity of $\mathcal{J}$). *Given are images $\mathcal{T}, \mathcal{R} \in \mathrm{Img}(\Omega)$, a distance functional $\mathcal{D}$, $\mathcal{S}^{\mathrm{hyper}}$, and $\mathcal{A}$ as in Definition 3. Let the assumptions (A1)–(A4) be satisfied.*

- *There exists a function $g : \Omega \times \mathbb{R}^3 \times \mathbb{R}^{3\times3} \times \mathbb{R}^{3\times3} \times (0, \infty] \to \mathbb{R}$ with*

$$\mathcal{J}[y] = \int_{\Omega} g(x, y, \nabla y, \mathrm{cof}\nabla y, \det \nabla y)\, dx$$

  that is continuously differentiable, convex in its last three arguments, and measurable in its first argument.
- *The objective functional $\mathcal{J}$ is weakly lower semicontinuous, i.e.,*

$$\liminf_{k\to\infty} \int_{\Omega} g(x, y^k, \nabla y^k, \mathrm{cof}\nabla y^k, \det \nabla y^k)\, dx \geq \int_{\Omega} g(x, y, \nabla y, H, v)\, dx$$

whenever $y^k \rightharpoonup y$ *in* $W^{1,2}$, $\quad \operatorname{cof}\nabla y^k \rightharpoonup H$ *in* L^4, $\quad \det \nabla y^k \rightharpoonup v$ *in* L^2.

Proof. Both results are proved separately.

First, $\mathcal{J}$ is rewritten as

$$\mathcal{J}[y] = \int_\Omega g_D(x, y, \nabla y, \operatorname{cof}\nabla y, \det \nabla y))\, dx + \int_\Omega \alpha_\ell(x)\, |\nabla y|^2 + \alpha_{\mathrm{a}}(x)\, \phi(\operatorname{cof}\nabla y) + \alpha_{\mathrm{v}}(x)\, \psi(\det \nabla y)\, dx$$

and combine the integrands to $g : \Omega \times \mathbb{R}^3 \times \mathbb{R}^{3\times 3} \times \mathbb{R}^{3\times 3} \times (0, \infty] \to \mathbb{R}$. By assumption on $\mathcal{D}$ and design of $\mathcal{S}^{\text{hyper}}$, the mapping $g(x, y, \cdot, \cdot, \cdot)$ is convex for fixed x and y. Furthermore, since the images and the functions ϕ and ψ are continuously differentiable, the integrand g is continuously differentiable. Since the images are measurable, g is measurable in x for fixed $(y, D, H, v) \in \mathbb{R}^3 \times \mathbb{R}^{3\times 3} \times \mathbb{R}^{3\times 3} \times (0, \infty]$.

For the second part, one notes that the weak convergence of y^k to y in $W^{1,2}$ implies strong convergence of y^k in L^2 due to the compact embedding of $W^{1,2} \subset\subset L^2$; see [25, Sect. 5.7]. Hence using the first part yields weak lower semicontinuity of $\mathcal{J}$ on X by applying Theorem 2.

Due to the coercivity of $\mathcal{J}$ shown in Lemma 1, there exists a bounded minimizing sequence $\{y^k, \operatorname{cof}\nabla y^k, \det \nabla y^k\}_k \subset X$, where X is the product space as in (13). Lemma 2 shows lower semicontinuity of $\mathcal{J}$, thus, the sequence converges to a minimizer (y, H, v) in X. In the next step, the relation between the components y, H and v is established. More specifically, a continuity results for the cofactor and determinant mappings for function in $W^{1,p}(\Omega, \mathbb{R}^3)$ is given.

Theorem 3 (Weak Continuity of Determinants). *Assume* $d < q < \infty$ *and* $y^k \rightharpoonup y$ *in* $W^{1,q}(\Omega, \mathbb{R}^d)$. *Then* $\det \nabla y^k \rightharpoonup \det \nabla y$ *in* $L^{q/d}(\Omega, \mathbb{R})$.

Proof. See Evans [25, p. 454] for a complete proof.

This result is expected as the minimizing sequences $\{y^k\}_k$ are bounded in the spaces $W^{1,q}(\Omega, \mathbb{R}^d)$ and $\{\det \nabla y^k\}_k$ as a polynomial of degree d can be bounded in $L^d(\Omega, \mathbb{R})$. Combining these two bounds yields the weak convergence of the sequence $\{\det \nabla y^k\}_k$ in $L^{q/d}(\Omega, \mathbb{R})$. This result is used in [58] to show existence of solutions to volume constrained image registration in $W^{1,p}$ and actually extended to $p \geq d$.

Using Theorem 3, one option to obtain the identification $v = \det \nabla y$ in L^2 is to tighten the measurability demands on y by, e.g., resorting to $W^{1,6}(\Omega, \mathbb{R}^3)$ similar to [60].

In hyperelastic image registration also the cofactor matrix is regularized by the area regularization functional. This can be used to improve the above result; see also [4, 18]. This allows to achieve an improved continuity result based on [18].

Theorem 4 (Weak Continuity of Cofactors and Determinants). *Let $\Omega \subset \mathbb{R}^3$ fulfill (A4). Then it holds*

$$\left.\begin{array}{rl} y^k \rightharpoonup y & \text{in } W^{1,2}(\Omega,\mathbb{R}^3), \\ \operatorname{cof}\nabla y^k \rightharpoonup H & \text{in } L^4(\Omega,\mathbb{R}^{3\times 3}), \\ \det \nabla y^k \rightharpoonup v & \text{in } L^2(\Omega,\mathbb{R}) \end{array}\right\} \Rightarrow \left\{\begin{array}{l} H = \operatorname{cof}\nabla y, \\ v = \det \nabla y. \end{array}\right.$$

Proof. This proof combines the proofs of Theorems 7.5-1 and 7.6-1 in [18] and is structured into two steps.

For ease of presentation, define $\Gamma := \operatorname{cof}\nabla y$.

Step 1: Weak continuity of the cofactor mapping. Without loss of generality, presentation is focussed on the top-left entry of the cofactor matrix, i.e., to show that $\Gamma_{11}(x) = H_{11}(x)$ for almost every $x \in \Omega$. The result extends straightforwardly to the remaining entries.

To begin with, it is assumed that $y \in \mathcal{C}^2(\Omega,\mathbb{R}^3)$. Then it holds

$$\Gamma_{11} = \partial_2 y_2 \partial_3 y_3 - \partial_2 y_3 \partial_3 y_2 = \partial_3(y_3 \partial_2 y_2) - \partial_2(y_3 \partial_3 y_2).$$

Since y is continuously differentiable and $\partial\Omega$ is smooth one can integrate by parts, see [25, Sect. C.2]. Thus for test functions $\zeta \in \mathcal{C}_0^\infty(\Omega,\mathbb{R})$ it holds

$$\int_\Omega \gamma_{11}\, \zeta \, dx = -\int_\Omega y_3 \partial_2 y_2 \partial_3 \zeta \, dx + \int_\Omega y_3 \partial_3 y_2 \partial_2 \zeta \, dx. \tag{27}$$

This equality is now extended to functions $y \in W^{1,2}(\Omega,\mathbb{R}^3)$ by the following approximation argument. Since $\mathcal{C}^\infty(\Omega,\mathbb{R}^3)$ is dense in $W^{1,2}(\Omega,\mathbb{R}^3)$ [25, Sect. 5.2], one can approximate y by a sequence of functions in $\mathcal{C}^2$ for which (27) holds. Note that for ζ being fixed both sides of (27) are continuous in the space $\mathcal{C}^2$ with the norm $\|\cdot\|_{W^{1,2}}$, since

$$\left|\int_\Omega \Gamma_{11}\zeta \, dx\right| \le \|\Gamma_{11}\|_{L^1} \, \|\zeta\|_{L^1} \le C_1(\zeta) \, \|y\|^2_{W^{1,2}} \tag{28}$$

and for $i, j, l = 1, 2, 3$,

$$\left|\int_\Omega y_i \partial_j y_l \partial_l \zeta \, dx\right| \le \|y_i\|_{L^1} \|\partial_j y_l\|_{L^1} \|\partial_l \zeta\|_{L^1} \le C_2(\zeta) \, \|y\|^2_{W^{1,2}}. \tag{29}$$

Hence, (27) holds also for $y \in W^{1,2}(\Omega,\mathbb{R}^3)$ and thus its right hand side can be used to prove the desired equality $\Gamma = H$. To this end, consider a fixed test function ζ. By Hölder's inequality [25, Sect. B.2], it holds that the bilinear mapping

$$B(\cdot,\cdot) : L^2(\Omega,\mathbb{R}) \times W^{1,2}(\Omega,\mathbb{R}) \to \mathbb{R} \text{ with } (f,g) \mapsto \int_\Omega f\ \partial_i g\ \partial_m \zeta\ dx$$

is continuous. Due to the compact embedding $W^{1,2} \subset\subset L^2$ [25, Sect. 5.1], it holds

$$y^k \rightharpoonup y \text{ in } W^{1,2}(\Omega,\mathbb{R}^3) \Rightarrow y^k \to y \text{ in } L^2(\Omega,\mathbb{R}^3).$$

Because of the continuity of the bilinear form B, the strong convergence of $\{y^k\}_k$, and the weak convergence of $\{\partial_j y_l^k\}_k$, Theorem 7.1-5 in [18, p.348f] can be used to obtain that for all $i,j,l,m \in \{1,2,3\}$

$$y^k \rightharpoonup y \text{ in } W^{1,2}(\Omega,\mathbb{R}^3) \Rightarrow \int_\Omega y_i^k\ \partial_j y_l^k\ \partial_m \zeta\ dx \to \int_\Omega y_i\ \partial_j y_l\ \partial_m \zeta\ dx,$$

which due to (27) implies $\int_\Omega (\mathrm{cof}\nabla y^k)_{11}\zeta dx \to \int_\Omega (\mathrm{cof}\nabla y)_{11}\zeta dx$. Note that H_{11} is by assumption the weak limit of $\{(\mathrm{cof}\nabla y^k)_{11}\}_k$ in $L^4(\Omega,\mathbb{R})$. Therefore, $(\mathrm{cof}\nabla y)_{11}(x) = H_{11}(x)$ for almost every x and thus both functions are equal in $L^4(\Omega,\mathbb{R})$. Repeating these arguments for the remaining matrix entries yields $\mathrm{cof}\nabla y = H$ in $L^4(\Omega,\mathbb{R}^{3\times 3})$.

Step 2: Weak continuity of the determinant mapping. It is now shown that under the above assumptions $\det \nabla y(x) = v(x)$ for almost every $x \in \Omega$.
Recall that $(\det A)\cdot I = A(\mathrm{cof}A)^\top$ and hence $\det A = \sum_{j=1}^3 a_{1j}(\mathrm{cof}A)_{1j}$. Starting with $y \in \mathcal{C}^2(\Omega,\mathbb{R}^3)$,

$$\det \nabla y = \sum_{j=1}^n (\partial_j y_1)\cdot \Gamma_{1j} = \sum_{j=1}^n \partial_j [y_1 \Gamma_{1j}], \tag{30}$$

where the last equality holds, since the cofactor is divergence free [25, Sect. 8.1]: $\sum_{j=1}^3 \partial_j \Gamma_{ij} = 0$ for all $i = 1,2,3$. Since y is continuously differentiable and $\partial\Omega$ is smooth, one can integrate by parts [25, Sect. C.2] and see from (30) that for all test functions $\zeta \in \mathcal{C}_0^\infty(\Omega,\mathbb{R})$

$$\int_\Omega \sum_{j=1}^3 (\partial_j y_1)\cdot \Gamma_{1j}\ \zeta\ dx = \int_\Omega \sum_{j=1}^3 \partial_j [y_1 \Gamma_{1j}]\ \zeta\ dx = -\int_\Omega \sum_{j=1}^n y_1\ \Gamma_{1j}\ \partial_j \zeta\ dx. \tag{31}$$

For fixed ζ, the mapping $y \mapsto \int_\Omega \sum_{j=1}^n \Gamma_{1j}\ \partial_j \zeta\ dx$ is continuous with respect to the $W^{1,2}(\Omega,\mathbb{R}^3)$ norm. By applying the same approximation argument as above, one sees that functions $y \in W^{1,2}(\Omega,\mathbb{R}^3)$ are divergence free in the following sense:

$$\int_\Omega \sum_{j=1}^n (\mathrm{cof}\nabla y)_{1j}\ \partial_j \zeta\ dx = 0 \text{ for all } \zeta \in \mathcal{C}_0^\infty(\Omega,\mathbb{R}). \tag{32}$$

Now (31) is extended to functions $y \in \{y \in W^{1,2}(\Omega, \mathbb{R}^3) : \text{cof}\nabla y \in L^4(\Omega, \mathbb{R}^{3\times 3})\}$ by showing that for all test functions $\zeta \in \mathcal{C}_0^\infty(\Omega, \mathbb{R})$,

$$\int_\Omega \sum_{j=1}^{3} (\partial_j y_1) \Gamma_{1j}\ \zeta\ dx = -\int_\Omega \sum_{j=1}^{3} y_1\ \Gamma_{1j}\ \partial_j \zeta dx. \tag{33}$$

To this end, note that for fixed ζ, both sides of this equation are bounded bilinear forms in y_1 and Γ and thus continuous. Therefore, it can be assumed that $y_1 \in \mathcal{C}^\infty(\Omega, \mathbb{R}^3)$. But then it follows that $(y_1\zeta) \in \mathcal{C}_0^\infty(\Omega, \mathbb{R})$ and thus

$$0 = \int_\Omega \sum_{j=1}^{3} \Gamma_{1j} \partial_j (y_1\zeta)\ dx = \int_\Omega \sum_{j=1}^{3} y_1\ \Gamma_{1j} \partial_j \zeta\ dx + \int_\Omega \sum_{j=1}^{3} (\partial_j y_1)\ \Gamma_{1j}\ \zeta\ dx,$$

which yields the desired extension of (31).

Given that $y^k \rightharpoonup y$ in $W^{1,2}(\Omega, \mathbb{R}^3)$ and $\text{cof}\nabla y^k \rightharpoonup \text{cof}\nabla y$ in $L^4(\Omega, \mathbb{R}^{3\times 3})$, one obtains the continuity result

$$\begin{aligned}\int_\Omega \det \nabla y^k\ \zeta\ dx &= -\int_\Omega \sum_{j=1}^{3} y_1^k (\text{cof}\nabla y^k)_{1j} \partial_j \zeta dx \\ &\to -\int_\Omega \sum_{j=1}^{n} y_1 (\text{cof}\nabla y)_{1j} \partial_j \zeta dx = \int_\Omega \det \nabla y\ \zeta\ dx\end{aligned}$$

as in the first part of the proof from the compact embedding of $W^{1,2}(\Omega, \mathbb{R}^3)$ into $L^2(\Omega, \mathbb{R}^3)$. Note that for fixed ζ both sides of (33) are bounded bilinear functions.

From $\det \nabla y^k \rightharpoonup v$ and $\det \nabla y^k \rightharpoonup \det \nabla y$ in $L^2(\Omega, \mathbb{R})$, it follows that $v = \det \nabla y$ almost everywhere, which concludes the proof.

This completes the ingredients for the proof of the main result in Theorem 1.

Proof (Proof of Theorem 1). By assumption, $\mathcal{J}$ is bounded from below. Thus, there exists a minimizing sequence $\{y^k\}_k$ such that

$$\lim_{k\to\infty} \mathcal{J}[y^k] = \inf_{y\in\mathcal{A}} \mathcal{J}[y^k] =: m.$$

Note that $\mathcal{J}$ is finite for $\text{Id}(x) = x$. Therefore, it can be assumed that $\{\mathcal{J}[y^k] : k \in \mathbb{N}\}$ is also bounded from above by a constant $M > 0$.

By Lemma 1, the sequence $\{(y^k, \text{cof}\nabla y^k, \det \nabla y^k)\}_k$ can be bounded in the Banach space X defined in (13) in terms of M. More precisely, there are constants $C > 0$ and $K \in \mathbb{R}$ such that for all k it holds

$$M \geq \mathcal{J}[y^k] \geq K + C(\|y^k\|^2_{W^{1,2}} + \|\mathrm{cof}\nabla y^k\|^4_{L^4} + \|\det \nabla y^k\|^2_{L^2}).$$

Since X is a reflexive space, there exists a subsequence – again denoted by $\{(y^k, \mathrm{cof}\nabla y^k, \det \nabla y^k)\}_k$ – that converges weakly to a $(y, H, v) \in X$.

Lemma 2 proofs $\mathcal{J}$ to be weak lower continuous. Thus

$$\liminf_{k\to\infty} \int_\Omega g(x, y^k, \nabla y^k, \mathrm{cof}\nabla y^k, \det \nabla y^k)\, dx \geq \int_\Omega g(x, y, \nabla y, H, v)\, dx.$$

Theorem 4 yields the identifications $H = \mathrm{cof}\nabla y$ and $v = \det \nabla y$.

The last step is to show that y is admissible. It remains to show that $\det \nabla y > 0$ almost everywhere. To this end, consider a fixed $\epsilon \in [0, 1)$ and the set

$$S_\epsilon := \{x \in \Omega \ : \ \det \nabla y(x) < \epsilon\}.$$

It then holds that

$$\begin{aligned}
\psi(\epsilon)\ \mathrm{vol}(S_\epsilon) &\leq \int_{S_\epsilon} \psi(\epsilon)\, dx \leq \int_{S_\epsilon} \psi(\det \nabla y)\, dx \\
&\leq \int_{S_\epsilon} g(x, y, \nabla y, \mathrm{cof}\nabla y, \det \nabla y)\, dx \\
&\leq \liminf_{k\to\infty} \int_{S_\epsilon} g(x, y^k, \nabla y^k, \mathrm{cof}\nabla y^k, \det \nabla y^k)\, dx \leq M.
\end{aligned}$$

Due to the growth condition $\psi(v) \to \infty$ for $v \to 0^+$, it follows that S_0 has zero volume. Thus, $\det \nabla y > 0$ almost everywhere in Ω and $y \in \mathcal{A}$, which completes the proof.

To summarize, the hyperelastic regularization functional $\mathcal{S}^{\mathrm{hyper}}$ guarantees the existence of optimal transformations for the unconstrained variational image registration problem. The proof of Theorem 1 illustrates difficulties due to the non-convexity of the objective functional and shows remedies. The main idea is to split y into Jacobian, cofactor, and determinant components and to show convergence of a minimizing sequence in a product space by proving coercivity and lower semicontinuity; see Lemmas 1 and 2. Subsequently, weak continuity of the cofactor and determinant mapping is used to undo this splitting; see Theorem 4. The provided proof is a special case of the existence theory of *polyconvex* functionals; see [4, 18, 20, 25]. In addition to providing existence of optimal transformations, the theory shows that a solution y satisfies $\det \nabla y > 0$ almost everywhere.

Instead of striving for the most generalized result, this section intends to provide insight into machinery of [4]. Therefore, four simplifying assumptions were made that are, however, uncritical in most practical applications.

4 Numerical Methods for Hyperelastic Image Registration

Generally, image registration problems cannot be solved analytically. Thus, numerical methods are used to approximate an optimal transformation. This section presents numerical methods for hyperelastic image registration based on a discretize-then-optimize approach. The first step is to properly discretize the variational problem (2). Typically, a coarse-to-fine sequence of finite dimensional optimization problems is generated, where all problems are linked by the underlying variational problem (2). Each problem can then be solved using standard optimization methods. Note that the coarse-to-fine strategy adds further regularization and additionally speeds up the computations.

In the hyperelastic setting, discretization is not straightforward. A challenge is to ensure one-to-one numerical solutions to the optimization problems, which are mandatory in most practical applications and guaranteed in the continuous function space setting. It is well known that a discrete object does not necessarily mimic all the properties of its continuous counterpart. For example, a function that is non-negative on a discrete set may not be non-negative everywhere. Thus, an approximated solution may not satisfy $\det \nabla y > 0$ almost everywhere, although the above existence theory guarantees this for the solutions of the continuous problem.

This section organizes as follows. Section "Discretizing the Determinant of the Jacobian" illustrates difficulties arising in the discretization of Jacobian determinants using finite difference schemes. As one way to avoid these difficulties, Galerkin finite element methods are presented in section "Galerkin Finite Element Discretization". The main idea is to compute a solution to the variational problem in the space of globally continuous and piecewise linear transformations. A key benefit is that for these transformations geometric quantities like area and volume can be computed exactly; see section "Galerkin Finite Element Discretization". Consequently, solutions of the discrete problem are guaranteed to be one-to-one. Finally, section "Multi-level Optimization Strategy" presents multi-level numerical optimization approaches that solve the finite dimensional optimization problems; see also [55].

Discretizing the Determinant of the Jacobian

The variational image registration problem (2) can be discretized using finite difference schemes. The underlying idea is to approximate the solution to the variational problem on a *grid*, which is essentially a collection of points. Many commonly used algorithms for image registration are based on these principles; see, e.g., [55] for an overview.

However, deriving finite difference discretization of operators such as the determinant of the Jacobian is not straightforward as illustrated by a simplified 2D example. For a transformation $y \in \mathcal{C}^2(\Omega, \mathbb{R}^2)$, the determinant of the Jacobian is

$$\det \nabla y(x) = \partial_1 y_1(x)\, \partial_2 y_2(x) - \partial_2 y_1(x)\, \partial_1 y_2(x). \tag{34}$$

Note that all entries of $\nabla y(x)$ are coupled, which causes difficulties for finite difference discretizations. This section analyzes this problem in detail and presents a finite volume technique that can be used to obtain a proper discretization.

Finite Differences in 1D

To start with the analysis, consider in a 1D setting discretization of the energy

$$\mathcal{S}[y] = \frac{1}{2} \int_{\Omega} (\partial y(x))^2 dx, \tag{35}$$

where $\Omega = [0, 1]$ is divided into m cells of width $h = 1/m$. The *cell-centered grid* $x^c \in \mathbb{R}^m$ and the *nodal grid* $x^n \in \mathbb{R}^{m+1}$ are defined by $x_i^c = (i - 0.5)h$ and $x_i^n = (i - 1)h$, respectively; see [55] for details. Assuming that $y \in \mathcal{C}^2(\Omega, \mathbb{R})$ and a *short central finite difference* yields

$$\partial y(x_i^c) = \frac{y(x_{i+1}^n) - y(x_i^n)}{h} + \mathcal{O}(h^2). \tag{36}$$

Note that the discretization is of second order for a price of a grid change. The function y is approximated on the nodal grid, whereas its derivative is approximated on the cell-centered grid. With the discrete partial differential operator $\partial_m^h : \mathbb{R}^{m+1} \to \mathbb{R}^n$,

$$\partial_m^h := \frac{1}{h} \begin{pmatrix} -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{pmatrix} \in \mathbb{R}^{m \times m+1}, \tag{37}$$

Equation (36) reads $\partial y(x^c) = \partial_m^h\, y(x^n) + \mathcal{O}(h^2)$. To avoid the grid change, one could also use the *long stencil* $\partial y(x_i^c) \approx (y(x_{i+1}^c) - y(x_{i-1}^c))/(2h)$. The price to be paid is the non-trivial null-space with high frequency components which are unseen by this discrete regularizer. While this might be tolerable on a fixed discretization level, it is a crucial drawback for multigrid schemes. Therefore, this discretization is not recommended. Finally, one could also think about a one-sided stencil, $\partial y(x_i^c) \approx (y(x_{i+1}^c) - y(x_i^c))/h$. Here, the price to be paid is that this approximation is only first order and this discretization is also not recommended. Thus, ∂_m^h is preferable for approximating derivatives of nodal quantities. Derivatives of cell-centered quantities can be approximated analogously on a subset of the nodal grid x^n using ∂_{m-1}^h.

Using a nodal discretization $y^n \approx y$ with $y_i^n \approx y(x_i^n)$ and the above discretization of the differential operator, a second-order approximation of (35) is obtained by using a midpoint quadrature rule:

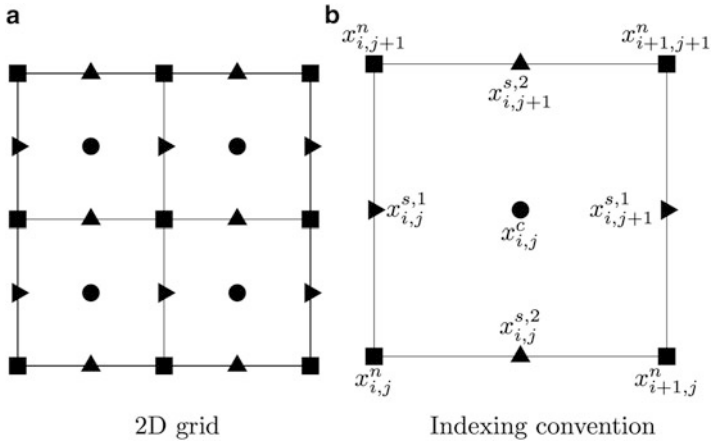


Fig. 4 Overview of grids and indexing convention in 2D as also used in [55]. A cell-centered grid x^c a nodal grid $x^n$and staggered grids $x^{s,1}$ and $x^{s,2}$ are depicted in (**a**). The indexing convention for the (i, j)th cell is illustrated in (**b**)

$$S[y^n] := \frac{h}{2}(y^n)^\top (\partial_m^h)^\top \partial_m^h \, y^n.$$

Note that both the finite difference operator and the quadrature rule are of second order which is essential for a consistent approximation.

Finite Differences in 2D

Already in 2D, the situation gets far more interesting. Let the rectangular domain $\Omega = [0, 1]^2$ be divided into (m, m) cells of uniform edge length $h = 1/m$. Similar as above, cell-centered and nodal grids are defined as

$$x^c = ((i - 0.5)h, (j - 0.5)h)_{i,j=1,\dots,m}$$
$$x^n = ((i - 1)h, (j - 1)h)_{i,j=1,\dots,m+1}.$$

See Fig. 4 for an overview of grids and indexing conventions. For ease of presentation of finite difference operators to higher dimensions, the concept of *Kronecker products* is convenient; see also [9].

Definition 4 (Kronecker Product). Given two matrices $A \in \mathbb{R}^{p_1 \times p_2}$ and $B \in \mathbb{R}^{q_1 \times q_2}$, the *Kronecker product* is defined by

$$A \otimes B = \begin{pmatrix} a_{1,1}B & \dots & a_{1,p_2}B \\ \vdots & \ddots & \vdots \\ a_{p_1,1}B & \dots & a_{p_1,p_2}B \end{pmatrix} \in \mathbb{R}^{p_1 q_1 \times p_2 q_2}.$$

As the following discussion illustrates, it is impossible to give a second order approximation to all components of the Jacobian matrix of $y = [y_1, y_2] \in \mathcal{C}^2(\Omega, \mathbb{R}^2)$ on a cell-centered grid by using short differences. To illustrate this, consider the first component y_1 of the transformation. Second order approximations to the partial derivatives of a nodal discretization $y_1^n \in \mathbb{R}^{(m+1)^2}$ can be obtained using the finite difference operators

$$\partial_{m,1}^h = I_{m+1} \otimes \partial_m^h \text{ and } \partial_{m,2}^h = \partial_m^h \otimes I_{m+1} \in \mathbb{R}^{m(m+1)\times(m+1)^2}.$$

In contrast to the 1D case, partial derivatives are not approximated on a cell-centered grid, but on *staggered grids*. Here, the first partial derivatives $\partial_{m,1}^h y_1^n$ and $\partial_{m,2}^h f^n$ are approximated at different locations

$$x^{s,1} = ((i-1)h, (j+.5)h)_{i=1,\dots,m+1, j=1,\dots,m},$$
$$x^{s,2} = ((i+.5)h, (j-1)h)_{i=1,\dots,m, j=1,\dots,m+1},$$

see also Fig. 4. As in the 1D case, grid changes are inevitable when using short differences. Furthermore, different partial derivatives are approximated at different positions. This also holds for other choices of discretization. For example, discretizing y_1 on the staggered grid $x^{s,1}$ and using the operators

$$\partial_{m,1}^h = I_m \otimes \partial_m^h \text{ and } \partial_{m,2}^h = \partial_{m-1}^h \otimes I_{m+1},$$

approximates $\partial_1 y_1$ on a cell-centered grid and $\partial_2 y_1$ on a subset of a nodal grid. Note that for all combinations of the above grids and finite difference operators the components of the gradient of y_1 are approximated at different locations. This is a major source of difficulty when discretizing regularization functionals based on coupled partial differential operators such as the determinant of the Jacobian (34).

Finite Volume Discretization

The above presentation indicates that second order approximations of the partial derivatives at the cell-centered grid with short differences are impossible. Averaging derivatives to the cell-centers is not a remedy as it annihilates oscillatory functions. One way to overcome this limitation is to approximate $\det \nabla y$ in a finite volume approach by measuring volume changes as it has been proposed [36]. In [36], the volume of a deformed cell V is approximated by the sum over of the volume of two triangles spanned by its vertices V_i, see also Fig. 5:

$$\begin{aligned}\int_V \det \nabla y(x) dx &= \int_{y(V)} dx \\ &= (V_3 - V_2) \times (V_2 - V_1)/2 + (V_4 - V_1) \times (V_3 - V_4)/2 + \mathcal{O}(h^2).\end{aligned} \tag{38}$$

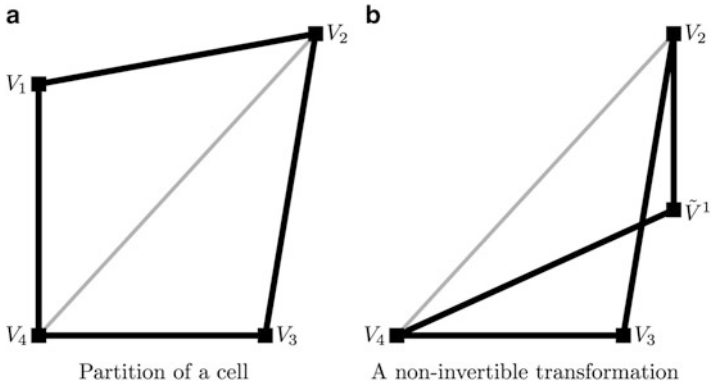


Fig. 5 Measuring the volume of a transformed cell as in [36]. The volume of the transformed cell is approximated by the sum of the volumes of the two triangles spanned by its vertices; see (**a**). In (**b**), a non-invertible transformation is applied to the cell. Note that although the volume of the triangle spanned by V_4, V_2 and $\tilde{V}^1$ is negative, the volume of V is positive. This motivates to individually control volume changes of the triangles in order to detect twists

Assuming $y \in \mathcal{C}^2(\Omega, \mathbb{R}^3)$, this approximation is second order; see also section "Images and Transformation". The first equality requires that y is invertible due to the transformation theorem, which is to be ensured by hyperelastic regularization. An advantage of this approach is that transformations that are not one-to-one can be detected. Note that for a non-invertible transformation and a sufficiently fine spacing of control points there is a sign change of the volume of a triangle; see Fig. 5. This approach was also used for hyperelastic image registration in [14].

Note that using this discretization, the constraint $\det \nabla y > 0$ is controlled only approximately and on a tetrahedral partition rather than point-wise. In order to ensure $\det \nabla y > 0$ almost everywhere, additional assumptions on the behavior of y between the control points are required. In [14] it is shown that based on the above discretization the volume constraint can be ensured almost everywhere for continuous and piecewise linear transformations. Motivated by this observation, the next section presents a Galerkin Finite Element approach that operates on this finite dimensional space.

Galerkin Finite Element Discretization

This section presents a *Galerkin finite element method* based on the general framework presented in [32]. The idea is to discretize the variational problem (2) with regularization functional $\mathcal{S}^{\text{hyper}}$ using globally continuous and piecewise linear transformations on a tetrahedral mesh. For these functions, the constraint $\det \nabla y > 0$ is fulfilled everywhere if it holds on each element. In accordance with the theory, our discrete solutions will share this important quality.

Let $V_1, \ldots, V_{n_V} \in \mathbb{R}^3$ denote vertices and $T_1, \ldots, T_{n_T}$ tetrahedra, where it is assumed that $\mathrm{vol}(T_i) > 0$ for all $i = 1, \ldots, n_T$. This builds a finite dimensional subspace

$$\mathcal{A}_h := \{y \in \mathcal{C}(\Omega, \mathbb{R}^3) : y\big|_{T_i} \in \Pi^1(T, \mathbb{R}^3) \text{ for } i = 1, \ldots, n_T\} \subset \mathcal{A}, \tag{39}$$

where h is related to n_V and denotes order of the approximation, Π_1 denotes the space of first order vector-valued polynomials. Standard nodal Lagrange that functions $b^1, \ldots, b^{n_V} : \Omega \to \mathbb{R}$ are used as a basis of $\mathcal{A}_h$. The coefficients $\mathbf{y}$ of $y^h \in \mathcal{A}_h$with respect to that basis are stored component-wise in a column vector of size $3n_V$, $\mathbf{y}_j = (\mathbf{y}_j^1, \mathbf{y}_j^2, \mathbf{y}_j^3)^\top$, $j = 1, \ldots, n_V$:

$$y^h(x) = \sum_{j=1}^{n_V} \mathbf{y}_j b^j(x), \tag{40}$$

where $\mathbf{y}_j = y^h(V_j)$ and analogously a discretization of a reference transformation $\mathbf{y}_{\mathrm{ref}}$ is obtained as $(\mathbf{y}_{\mathrm{ref}})_j = y_{\mathrm{ref}}(V_j)$. The discretized hyperelastic registration problem then reads

$$J : \mathbb{R}^{3n_V} \to \mathbb{R}^+, \quad J(\mathbf{y}) = D(\mathbf{y}) + \alpha \mathcal{S}^{\mathrm{hyper}}(\mathbf{y} - \mathbf{y}_{\mathrm{ref}}). \tag{41}$$

The discretization of the distance and hyperelastic regularization functional are derived as follows.

Let $I_k \in \mathbb{R}^{k \times k}$ denote the identity matrix, $1_k \in \mathbb{R}^k$ the vector of all ones, $\otimes$ the Kronecker product, and $\odot$ the Hadamard product. The entries of the Jacobian matrix ∇y^h are piecewise constant on each triangle. The nine components of the Jacobi matrix are stored in a column vector $(B\mathbf{y}) \in \mathbb{R}^{9n_T}$, where B is the discrete vector gradient operator formed essentially from the discrete partial derivatives $\partial_k^h \in \mathbb{R}^{n_T \times n_V}$ with $(\partial_k^h)_{i,j} = \partial_k b^i(V_j)$:

$$B = I_3 \otimes \nabla^h, \text{ with } (\nabla^h)^\top = \left((\partial_1^h)^\top, (\partial_2^h)^\top, (\partial_3^h)^\top\right). \tag{42}$$

Interpolation from nodes to barycenters of the tetrahedra is accomplished by applying the averaging matrix

$$\mathbf{A} = I_3 \otimes A \in \mathbb{R}^{n_T \times n_V} \text{ with } A_{i,j} = \begin{cases} 1/4 & \text{if } V_j \text{ is node of } T_i \\ 0 & \text{otherwise} \end{cases}. \tag{43}$$

Assembling the volumes of the tetrahedra in

$$\mathbf{v} \in \mathbb{R}^{n_T}, \text{ with } \mathbf{v}_i := \mathrm{vol}(T_i) \text{ and setting } \mathbf{V} := \mathrm{diag}(\mathbf{v}), \tag{44}$$

the SSD distance functional (5) is approximated by using a midpoint quadrature rule

$$D(\mathbf{y}) \approx \tfrac{1}{2}\mathrm{res}(\mathbf{y})^\top \mathbf{V}\mathrm{res}(\mathbf{y}), \text{ where } \mathrm{res}(\mathbf{y}) = \mathcal{T}(\mathbf{A}\mathbf{y}) - \mathcal{R}(\mathbf{x}). \tag{45}$$

Since the Jacobian matrix of $y^h \in \mathcal{A}_h$ is piecewise constant, the hyperelastic regularization functional is evaluated exactly by

$$\begin{aligned}\mathcal{S}^{\text{hyper}}(\mathbf{y}) =& \tfrac{\alpha_\ell}{2}(\mathbf{y}-\mathbf{y}_{\text{ref}})^\top B^\top (I_9 \otimes \mathbf{V}) B\ (\mathbf{y}-\mathbf{y}_{\text{ref}}) \\ &+ \alpha_{\text{a}}\mathbf{v}^\top \phi(\mathrm{cof} B\mathbf{y}) + \alpha_{\text{v}}\mathbf{v}^\top \psi(\det B\mathbf{y}).\end{aligned} \tag{46}$$

In line with the organization of the gradient, the entries of the cofactor matrix are stored in a column-vector of length $9n_T$. The first derivatives of the discretized objective function is

$$dJ(\mathbf{y}) = d\mathcal{D}(\mathbf{y}) + d\mathcal{S}^{\text{length}}(\mathbf{y}) + d\mathcal{S}^{\text{area}}(\mathbf{y}) + d\mathcal{S}^{\text{vol}}(\mathbf{y}). \tag{47}$$

Using the chain rule yields

$$d\mathcal{D}(\mathbf{y}) = \mathrm{res}(\mathbf{y})^\top \mathbf{V}(\nabla \mathcal{T}(\mathbf{A}\mathbf{y}))\mathbf{A}, \tag{48}$$

$$d\mathcal{S}^{\text{length}}(\mathbf{y}) = \alpha_\ell\ (\mathbf{y}-\mathbf{y}_{\text{ref}})^\top B^\top (I_9 \otimes \mathbf{V}) B, \tag{49}$$

$$d\mathcal{S}^{\text{area}}(\mathbf{y}) = \alpha_{\text{a}}\ ((1_9 \otimes \mathbf{v}) \odot \phi'(\mathrm{cof} B\mathbf{y}))^\top\ d\ \mathrm{cof} B\mathbf{y}, \tag{50}$$

$$d\mathcal{S}^{\text{vol}}(\mathbf{y}) = \alpha_{\text{v}}\ \mathbf{v}^\top\ \psi'(\det B\mathbf{y})\ d\ \det B\mathbf{y}. \tag{51}$$

Setting $D_i^j = \mathrm{diag}(\partial_i \mathbf{y}^j) \in \mathbb{R}^{n_T \times n_T}$, one gets

$$d\,\mathrm{cof} B\mathbf{y} = \begin{pmatrix} & & & & & D_3^3 & -D_2^3 & & -D_3^2 & D_2^2 \\ & & & & & D_3^3 & & -D_2^3 & & -D_3^1 & D_1^2 \\ & & & & & D_2^3 & -D_1^3 & & -D_2^2 & D_1^2 & \\ & D_3^3 & -D_2^3 & & & & & & & -D_3^1 & D_2^1 \\ D_3^3 & & -D_1^3 & & & & & & -D_3^1 & & D_1^1 \\ D_2^3 & -D_1^3 & & & & & & & -D_1^2 & D_1^1 & \\ & D_3^2 & -D_2^2 & & & -D_3^1 & D_2^1 & & & & \\ D_3^2 & & -D_1^2 & -D_3^1 & & & D_1^1 & & & & \\ D_2^2 & -D_1^2 & & D_2^1 & D_1^1 & & & & & & \end{pmatrix} B \in \mathbb{R}^{9n_T \times 3n_V}. \tag{52}$$

With the abbreviation $C_i^j = \mathrm{diag}((\mathrm{cof} B\mathbf{y})_{i,j}) \in \mathbb{R}^{n_T \times n_T}$, the derivative of the Jacobi determinant can be compactly written as

$$d\ \det B\mathbf{y} = \begin{pmatrix} C_1^1, & C_1^2, & C_1^3, & C_2^1, & C_2^2, & C_2^3, & C_3^1, & C_3^2, & C_3^3 \end{pmatrix}\ B \in \mathbb{R}^{n_T \times 3n_V}. \tag{53}$$

Following the guidelines of [55], the Hessian of the objective functional is approximated to ensure positive semi-definiteness and avoid computations of second derivatives of the (generally noisy) template image as follows

$$H(\mathbf{y}) \approx d_2 D(\mathbf{y}) + d_2 \mathcal{S}^{\text{length}} + d_2 \mathcal{S}^{\text{area}}(\mathbf{y}) + d_2 \mathcal{S}^{\text{vol}}(\mathbf{y}) \tag{54}$$

with the summands

$$d_2 D(\mathbf{y}) = \mathbf{A}^\top (\nabla \mathcal{T}(\mathbf{A}\mathbf{y}))^\top \mathbf{V} \nabla \mathcal{T}(\mathbf{A}\mathbf{y}) \mathbf{A}, \tag{55}$$

$$d_2 \mathcal{S}^{\text{length}} = \alpha_\ell B^\top (I_9 \otimes \mathbf{V}) B, \tag{56}$$

$$d_2 \mathcal{S}^{\text{area}}(\mathbf{y}) = \alpha_{\text{a}} (d \operatorname{cof} B\mathbf{y})^\top \left(\operatorname{diag}(1_9 \otimes \mathbf{v})\, \phi''(\operatorname{cof} B\mathbf{y})\right) d \operatorname{cof} B\mathbf{y}, \tag{57}$$

$$d_2 \mathcal{S}^{\text{vol}}(\mathbf{y}) = \alpha_{\text{v}} (d \det B\mathbf{y})^\top \operatorname{diag}(\mathbf{v} \odot \psi''(\det B\mathbf{y}))\, d \det B\mathbf{y}. \tag{58}$$

Note that for this distance functional the dependency of the first summand on $\mathbf{y}$ is only low order and the second summand is constant with respect to $\mathbf{y}$. In contrast, the Hessian of the area and volume regularization functionals strongly depend on $\mathbf{y}$.

Multi-level Optimization Strategy

Using the above presented Galerkin method translates the variational image registration problem into a finite dimensional optimization problem. Solving the discrete problem is not straightforward as it is typically large, nonlinear and non-convex. However, state-of-the-art tools from numerical optimization can be used to obtain accurate and efficient schemes. This section describes a multi-level strategy using Gauss–Newton optimization.

Multi-level Idea

In image registration, challenges arise from the non-convex dependence of the objective functional on the transformation to be determined. Thus, many local minima have to be expected; for an example, see [55, p. 112]. A typical solution strategy is to approximately solve the variational problem (2) on a coarse-to-fine hierarchy of discretizations often referred to as *multi-level* strategy [55] or *cascadic multigrid* [7]. The key motivation is to reduce the risk of being trapped in a local minimum and to obtain good starting guesses for the correction steps on finer discretization levels. A positive side effect is a reduction of computational costs.

Generically, a nested series of finite dimensional spaces $\mathcal{A}_{h_i}$ with

$$\mathcal{A}_{h_1} \subset \mathcal{A}_{h_2} \subset \cdots \subset \mathcal{A}$$

is constructed, where h_i refers to the approximation order. The idea then is to compute a minimizer y^1 of $\mathcal{J}$ in $\mathcal{A}_{h_1}$ first. This is relatively simple, as it is assumed that the space $\mathcal{A}_{h_1}$ is smooth and the problem is of small size. Note that y^1 is in S_{h_2} and can thus be used as starting guess. In practice, only the coefficients of the basis functions in S_{h_2} at the new vertices need to be computed. For Lagrange basis functions, weights are easily obtained by evaluating y^1 at the nodes of the mesh

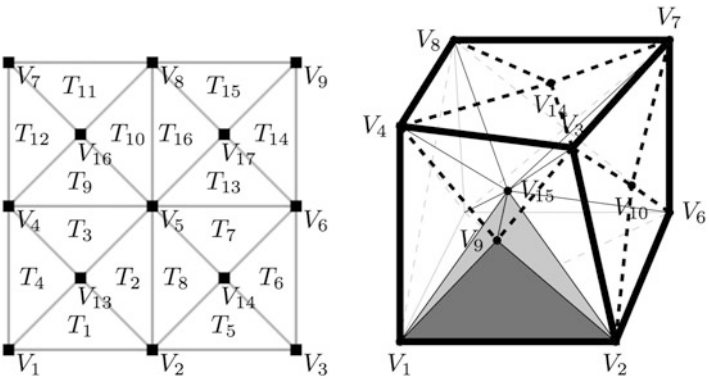


Fig. 6 Subdivision used for generating nested families of structured meshes as in [14]. *Left*: triangulation of an $m = [2, 2]$ grid. *Right*: tetrahedral subdivision of a voxel

of S_{h_2}. With this excellent starting guess, computing a solution in $\mathcal{A}_{h_2}$ is relatively simple, particularly with a second order schemes. This procedure is repeated until the finest level is reached.

There are many ways to construct a nested sequence of triangular meshes. A simple but powerful method is to use nested rectangular meshes and decompose each cell into tetrahedra as depicted in Fig. 6. However, more efficient strategies using local refinements based on error estimates exist; see, e.g., [39].

Commonly, the representation of the images are simplified on a coarse level by reducing the data; see [55]. Thereby, only essential image features are considered and consequently the danger of being trapped in a local minima is reduced further.

Gauss Newton Optimization

On each level, a Gauss–Newton scheme [57] can be used to compute a minimizer of the discretized objective function J in (41). Until convergence, the objective function $J(\mathbf{y})$, its gradient $dJ(\mathbf{y})$, and an approximation $H(\mathbf{y})$ to its Hessian are computed. The search direction v is then obtained by solving the linear system

$$H(\mathbf{y})\, v = -dJ(\mathbf{y}). \tag{59}$$

As outlined in section "Galerkin Finite Element Discretization", a symmetric and positive definite approximation of the Hessian is computed. Solving the above linear system is one of the most computationally expensive parts of the scheme. Recall (54) that the Hessian consists of four summands

$$H(\mathbf{y}) \approx d_2 D(\mathbf{y}) + d_2 \mathcal{S}^{\text{length}} + d_2 \mathcal{S}^{\text{area}}(\mathbf{y}) + d_2 \mathcal{S}^{\text{vol}}(\mathbf{y}).$$

The Hessian of the length regularizer $\mathcal{S}^{\text{length}}$ is simply a discrete Laplacian for each component of the transformation and does not depend on $\mathbf{y}$ unlike the other

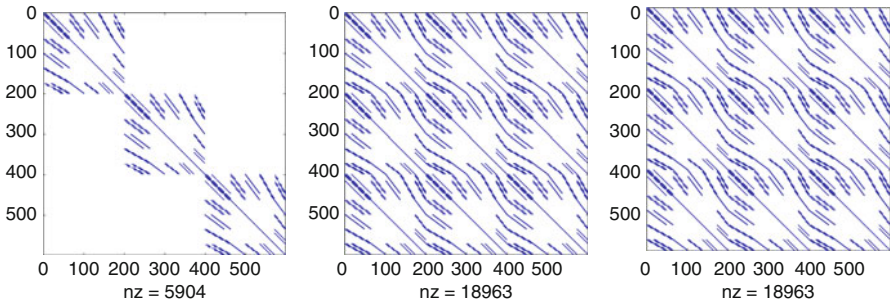


Fig. 7 Nonzero patterns of the Hessians of length (*left*), area (*middle*), and volume regularization functionals (*right*) for a tetrahedra mesh based on (3,3,3) cells. The mesh has 199 nodes and a number of nonzero elements in the Hessians are 5,904, 18,963, and 18,963, respectively

summands, which need to be updated in each iteration. Recall that a tight coupling between the components of the transformation is introduced by the area and volume regularizer and that the effective weight of $d^2\mathcal{S}^{\text{vol}}$ depends on the current iterate. A visualization of the nonzero patterns of the four blocks is presented in Fig. 7.

Since H is symmetric and positive semi-definite by design, a conjugate gradient (CG) method [44] can be used to approximate v. Conjugate gradient methods become even more efficient when being combined with preconditioning; see, e.g., [8]. Suppose there is an approximation $C \in \mathbb{R}^{N\times N}$ to H that is positive definite and easily invertible. Given a starting guess $v_0 \in \mathbb{R}^N$ consider the iteration

$$v_1 = v_0 - \alpha C^{-1}(H[y]v_0 + d\mathcal{J}[y])$$

with some step length α. Using $C = H$ and $\alpha = 1$, this would yield the result in one iteration. However, it is equivalent to directly solving (59). Thus, the idea is to use a "sufficient" approximation of C to H that is considerably easier to invert than H itself.

There are many preconditioning techniques for conjugate gradient methods. A computational attractive option is *Jacobi preconditioning*, i.e., $C = D$, where $D = \text{diag}(H)$. More expensive but also more effective preconditioning is offered by *Gauss–Seidel preconditioning*, i.e., $C = (D + L)D^{-1}(D + L^\top)$, where L is the lower triangular matrix of H and D the diagonal of H; see [8, Ch.4] for more details.

The step length is determined by a standard Armijo line search in combination with a backtracking that ensures $\det \nabla y > 0$; see [57]. Standard stopping rules based on the value of the objective functional, the norm of the gradient, and the norm of the update can be used as discussed in detail in [55].

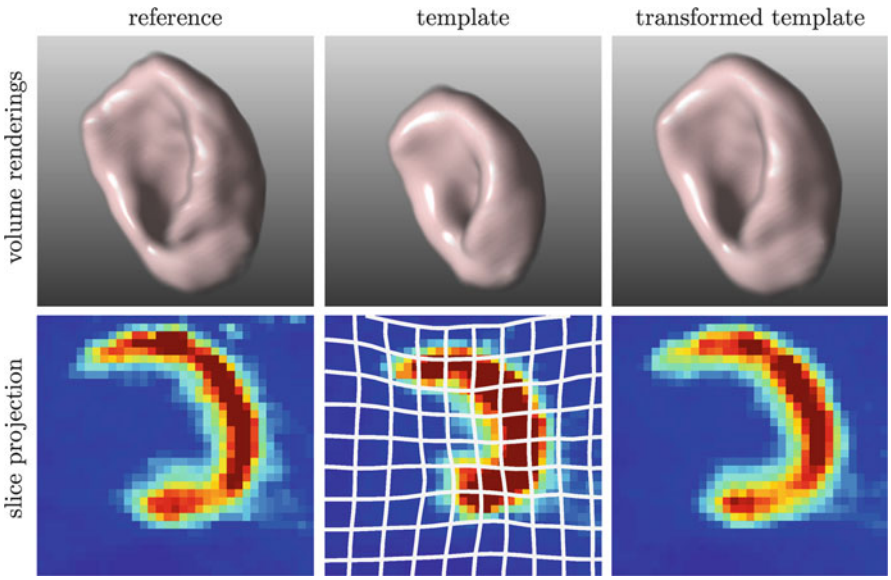


Fig. 8 3D Registration result for diastolic and systolic PET data of a human heart. The *first row* shows volume renderings of reference, template, and transformed template image. The *second row* shows slice projections of the image data and the estimated transformation. Image data provided by Fabian Gigengack, European Institute of Molecular Imaging, University of Münster, Germany

5 Applications of Hyperelastic Image Registration

This section outlines the potential of hyperelastic image registration for real-world applications. Exemplarily, two registration problems from medical imaging are considered. In both applications mass-preserving distance functionals are required and hyperelastic regularization is used to yield mathematically sound formulations.

Motion Correction of Cardiac PET

Positron Emission Tomography (PET) can provide useful information about the metabolism of the human heart, which is useful for the diagnosis of coronary artery diseases. To this end, a radioactive glucose tracer is administered and emission events are recorded by a detector ring during its decay. Finally, a tomographic image representing the spatial distribution of the tracer is reconstructed. Typically, a large fraction of the tracer is metabolized by the left ventricular muscle of the heart, which is thus visible in the images; see Fig. 8.

A key problem in PET reconstruction is the severe degradation of the images due to respiratory and cardiac motion during image acquisition. In principle, the more emission events are recorded, the better the expected image quality. Thus,

acquisition times are typically in the order of minutes and motion is inevitable. The so-called gating techniques are typically used to compensate for these motion artifacts [21]. The idea is that recorded emission events are grouped into a number of gates, which relate to particular phases in the respiratory and cardiac cycle. For each gate, a reconstruction is computed which shows less motion blur, but is also based on fewer counts and consequently of degraded quality as signal to noise has been reduced. To take full advantage of all measurements, the individual reconstructions are aligned using image registration and finally fused [30, 61].

Mass-Preservation

PET images represent the distribution of the density of the radioactive tracer, in this case, the distribution of a glucose tracer measured 1 h after injection. Therefore, it can be assumed the amount of tracer per given tissue unit to be constant. Thus the mass-preserving distance measure, as defined in section "Distance Functionals", is used

$$\mathcal{D}^{\mathrm{MP}}[y] = \mathcal{D}^{\mathrm{MP}}[\mathcal{T}, \mathcal{R}; y] = \frac{1}{2} \int_{\Omega} (\mathcal{T}[y] \cdot \det \nabla y - \mathcal{R})^2 dx.$$

As mentioned in section "Distance Functionals", the mass-preserving distance functional $\mathcal{D}^{\mathrm{MP}}$ is not convex in ∇y and mass-preservation requires that transformations satisfy $\det \nabla y > 0$. Also, large strains for cardiac motion correction are expected and images will be of relatively poor quality when using fine gating schemes. Thus, hyperelastic regularization is used since Theorem 1 guarantees existence of minimizers with positive Jacobian determinants.

The mass-preserving distance functional is discretized following the guidelines of section "Galerkin Finite Element Discretization". As for the SSD distance, a midpoint quadrature rule is employed and computation of the Jacobian determinant and its derivatives are reused.

Registration Results and Impact on Image Quality

Registration results for a 3D PET data set of a human heart are visualized in Fig. 8. More precisely, Fig. 8 only presents results for the registration of diastolic and systolic phases which is the most difficult scenario. Data courtesy by Fabian Gigengack, European Institute for Molecular Imaging, University of Münster, Germany. It turns out that the mass-preserving hyperelastic registration approach accurately deforms the contracted systolic gate such that is very similar to the diastolic gate. The smooth and invertible transformation is visualized using a slice projection.

The image quality that can dramatically improved by registration and subsequent averaging of the transformed images as shown in Fig. 9. Here, a dual gating into five respiratory gates each being divided into five cardiac gates is performed, yielding 25 image volumes. Each single gate contains only a small fraction of the signal, however, is less affected by motion; cf. Fig. 9a. Reconstruction without registration yields a smoother image degraded by motion blurr; cf. Fig. 9b. A smooth and

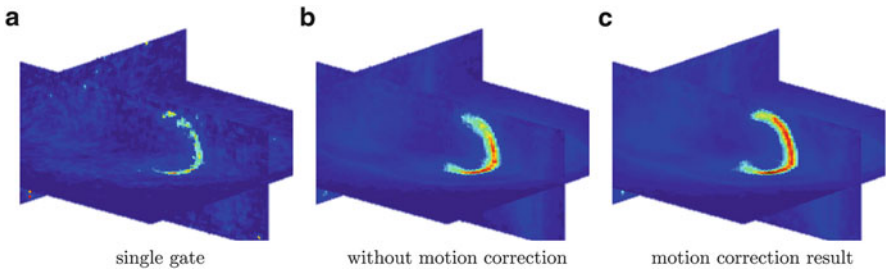


Fig. 9 Motion Correction result for PET data of a human heart. Slice projections of a single gate (*left*) a blurred reconstruction without motion correction (*center*) and motion-corrected reconstruction (*right*) are shown. Image data provided by Fabian Gigengack, European Institute of Molecular Imaging, University of Münster, Germany

motion-corrected reconstruction is obtained by following the pipeline suggested in [61] and averaging the aligned images. Here a much sharper and clearer image can be reconstructed; cf. Fig. 9c.

Summary and Further Literature

Mass-preserving hyperelastic image registration reduces motion artifacts in cardiac PET. The transformation model respects the density property of PET. Hyperelastic regularization is used to ensure existence of solutions that satisfy the mass-preserving constraint. The mass-preserving transformation model was validated in [30] and a pipeline for practical motion correction is suggested in [61].

Susceptibility Artefact Correction of Echo-Planar MRI

Echo Planar Imaging (EPI) is a commonly available ultrafast MRI acquisition technique; see [66]. It is routinely used for key investigation techniques in modern neuroscience such as diffusion tensor imaging (DTI) or functional MRI (fMRI).

While offering considerable reduction of acquisition time, a drawback of EPI is its high sensitivity against small perturbations of the magnetic field. Field inhomogeneities are inevitably caused by different magnetic susceptibilities of soft tissue, bone, and air and thus present in any practical setting. Inhomogeneities result in geometrical distortions and intensity modulations that complicate the interpretation of EPI data and their fusion with anatomical T1- or T2-weighted MR images obtained using conventional acquisition techniques with negligible distortions.

This section summarized the tailored variational image registration method for the correction of distortions presented in [62]. The method can be seen as a special case of mass-preservation hyperelastic registration for transformations that are restricted along one a priori known spatial direction.

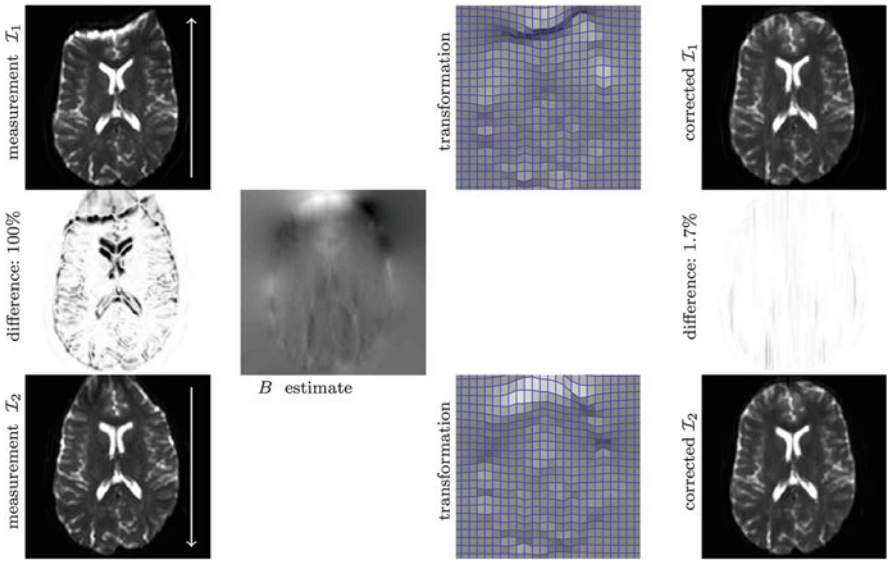


Fig. 10 Illustration of the Reversed Gradient Approach [15]. The initial measurements (*first column*) are distorted in opposite directions, depicted as *white arrows*. The estimated field inhomogeneity B (*second column*) gives rise to two geometric transformations (*third column*) that can be used to approximately correct the data (*right column*). Data courtesy by Harald Kugel, Department for Radiology, University of Münster, Germany

Reversed Gradient Method

The *Reversed Gradient Method* was firstly described by Chang and Fitzpatrick in 1992 [15] to correct for distortions due to field inhomogeneities. They derived a physical model for the distortions, which shows that measurements are distorted along one a priori known spatial direction. Interestingly, the direction of distortion v is not only known, but can also be controlled by parameters settings on the scanner. It is thus possible to acquire a pair of images, denoted by $\mathcal{I}_1$ and $\mathcal{I}_2$, that are oppositely affected by distortions; see also Fig. 10.

The Variational Problem

The measurements $\mathcal{I}_1$ and $\mathcal{I}_2$ represent density distributions of protons. Therefore, there is a mass-preserving property [15] is assumed. Due to field inhomogeneities, the signal is wrongly localized along the distortion direction v and the image intensities are modulated depending on the determinant of the Jacobian of the transformation (Fig. 11). Chang and Fitzpatrick describe the relation of the distorted measurements to the undistorted image $\mathcal{I}$ by

$$\mathrm{I}(x) = \mathcal{I}_1(x + vB(x))(1 + \partial_v B) = \mathcal{I}_2(x - vB(x))(1 + \partial_v B), \tag{60}$$

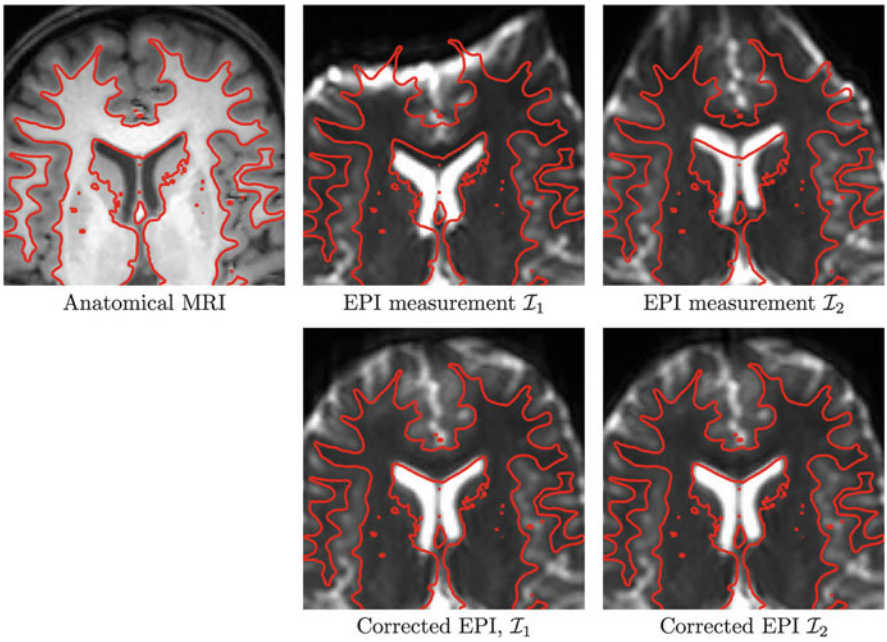


Fig. 11 Improved correspondence of corrected EPI measurements to anatomical MRI. Data courtesy by Harald Kugel, Department for Radiology, University of Münster, Germany

where $B : \Omega \to \mathbb{R}$ denotes the field inhomogeneity that needs to be measured or estimated numerically. Note that for distortion directions $\pm v$, there are two transformations $y^{\pm} = x \pm vB(x)$ and hence $\det \nabla y = 1 \pm \partial_v B$.

Considerable progress has been made in the numerical treatment of the reversed gradient method in the last decade; see, for instance [2,47,62,64]. These numerical schemes are based on a tailored mass-preserving distance functional

$$\mathcal{D}^{\text{EPI}}[B] = \frac{1}{2} \int_{\Omega} (\mathcal{I}_1(x + vB(x))(1 + \partial_v B) - \mathcal{I}_2(x - vB(x))(1 - \partial_v B))^2 \, dx.$$

A notable difference to image registration problems is that there is no template and reference image relationship. The displacement is applied to images, but in opposite directions. Further, intensity modulations are applied to ensure mass-preservation.

Regularization Because of the mass-preservation component invertibility of the geometrical transformations in (60) is required as has been already stressed in [15]. However, simplifications of the hyperelastic regularization functional $\mathcal{S}^{\text{hyper}}$ arise, since displacements are restricted in one direction. Thus the field inhomogeneity B needs to satisfy

$$0 < (1 + \partial_v B) \text{ and } 0 < (1 - \partial_v B) \Leftrightarrow -1 < \partial_v B < 1. \tag{61}$$

This motivates using the following regularizer

$$\mathcal{S}^{\text{EPI}}[B] = \frac{\alpha}{2} \int_\Omega \|\nabla B\|^2 dx + \beta \int_\Omega \psi(\partial_v B) dx.$$

The first summand is a diffusion term that ensures smoothness as commonly used and suggested by the forward model; see [22]. The convex penalty function

$$\psi(v) = \frac{v^4}{(v-1)(v+1)}$$

has singularities at ± 1 and ensures $\partial_v B \in L^2$. This is essential to ensure positivity of the intensity modulations. It is shown in [62] that this regularizer ensures existence of solutions and that solutions satisfy (61).

Numerical Implementation Numerical discretization of the reversed gradient can be done using a nodal grid for the field inhomogeneity and a finite volume approach. Intensity modulations are thus approximated by monitoring volume changes. Note that no coupling between the components of the Jacobian matrix is present since displacements are restricted along $\pm v$. A Matlab implementation is available as an extension to the FAIR toolbox [55] and integrated into the toolbox Statistical Parametric Mapping (SPM) [63].

Correction Results

Figure 10 illustrates the correction method for EPI data of the brain of a healthy volunteer; see [62] for more results. Data courtesy by Harald Kugel, Department for Radiology, University of Münster, Germany. The distance between the images $\mathcal{I}_1$ and $\mathcal{I}_2$ with respect to $\mathcal{D}^{\text{EPI}}$ is reduced considerably by the estimated transformation.

Figure 10 shows the improvement of spatial correspondence between the distorted EPI data and an anatomical T1-weighted MRI which can be assumed to be free of distortions. An axial slice zoomed into the frontal brain region is shown. Superimposed contour lines representing a white matter segmentation obtained from the anatomical image data are depicted in all subplots. Geometrical mismatch of the uncorrected EPI measurements and the anatomical data is most pronounced in frontal regions and around the Corpus Callosum. After correction, the geometrical correspondence improves considerably.

6 Conclusion

Image registration is an essential task in a variety of areas involving imaging techniques. This chapter presented a comprehensive overview of mathematical techniques used for nonlinear image registration. Emphasis was on regularization

techniques that ensure a mathematically sound formulation of the problem, allow stable and fast numerical solution, and favor solutions that are realistic for the application in mind.

Starting out from one of the most commonly used linear elastic model [10, 55], its limitations and extensions to nonlinear regularization functionals based on the theory of hyperelastic materials were discussed. A detailed overview of the available theoretical results was given. Insight into the existence theory of hyperelastic image registration problems was given and a state-of-the-art numerical scheme is presented. Finally, the potential of hyperelastic image registration for real-life medical imaging applications was outlined in two case studies.

Cross-References

- ▸ Large-scale Inverse Problems in Imaging
- ▸ Mathematical Methods in PET and SPECT Imaging
- ▸ Optical Flow
- ▸ Regularization Methods for Ill-Posed Problems

References

1. Ambrosio, L.: Lecture notes on optimal transport problems. In: Mathematical Aspects of Evolving Interfaces. Lecture Notes in Mathematics, vol. 1812, pp. 1–52. Springer, Berlin/Heidelberg (2003)
2. Andersson, J.L.R., Skare, S., Ashburner, J.: How to correct susceptibility distortions in spin-echo echo-planar images: application to diffusion tensor imaging. NeuroImage **20**(2), 870–888 (2003)
3. Ashburner, J.: A fast diffeomorphic image registration algorithm. NeuroImage **38**(1), 95–113 (2007)
4. Ball, J.M.: Convexity conditions and existence theorems in nonlinear elasticity. Arch. Ration. Mech. Anal. **63**(4), 337–403 (1976)
5. Beg, M., Miller, M.I., Trouvé, A.: Computing large deformation metric mappings via geodesic flows of diffeomorphisms. Int. J. Comput. Vis. **61**(2), 139–157 (2005)
6. Benamou, J.-D., Brenier, Y.: A computational fluid mechanics solution to the Monge–Kantorovich mass transfer problem. Numer. Math. **84**, 375–393 (2000)
7. Bornemann, F.A., Deuflhard, P.: The cascadic multigrid method for elliptic problems. Numer. Math. **75**, 125–152 (1996)
8. Braess, D.: Finite Elements. Theory, Fast Solvers, and Applications in Elasticity Theory, 2nd edn. Cambridge University Press, New York (2001)
9. Brewer, J.W.: Kronecker products and matrix calculus in system theory. IEEE Trans. Circuits Syst. **25**(9), 772–781 (1978)
10. Broit, C.: Optimal Registration of Deformed Images. Ph.D. thesis, University of Pennsylvania (1981)
11. Brown, L.G.: A survey of image registration techniques. ACM Comput. Surv. (CSUR) **24**(4), 325–376 (1992)
12. Bruhn, A., Weickert, J., Schnörr, C.: Lucas/Kanade meets Horn/Schunck: combining local and global optic flow methods. Int. J. Comput. Vis. **61**(3), 211–231 (2005)

13. Brune, C.: 4D Imaging in Tomography and Optical Nanoscopy. Ph.D. thesis, University of Münster (2010)
14. Burger, M., Modersitzki, J., Ruthotto, L.: A hyperelastic regularization energy for image registration. SIAM J. Sci. Comput. **35**, B132–B148 (2013)
15. Chang, H., Fitzpatrick, J.M.: A technique for accurate magnetic resonance imaging in the presence of field inhomogeneities. IEEE Trans. Med. Imaging **11**(3), 319–329 (1992)
16. Christensen, G.E.: Deformable Shape Models for Anatomy. Ph.D. thesis, Washington University (1994)
17. Christensen, G.E., Johnson, H.: Consistent image registration. IEEE Trans. Med. Imaging **20**(7), 568–582 (2001)
18. Ciarlet, P.G.: Mathematical Elasticity: Three Dimensional Elasticity. North Holland, Amsterdam (1988)
19. Collignon, A., Maes, F., Delaere, D., Vandermeulen, D., Suetens, P., Marchal, G.: Automated multi-modality image registration based on information theory. Inform. Process. Med. Imaging **14**(6), 263–274 (1995)
20. Dacorogna, B.: Direct Methods in the Calculus of Variations, 2nd edn. Springer, New York (2008)
21. Dawood, M., Stegger, L., Wübbeling, F., Schäfers, M., Schober, O., Schäfers, K.P.: List mode-driven cardiac and respiratory gating in PET. J. Nucl. Med. **50**(5), 674–681 (2009)
22. De Munck, J.C., Bhagwandien, R., Muller, S.H., Verster, F.C., Van Herk, M.B.: The computation of MR image distortions caused by tissue susceptibility using the boundary element method. IEEE Trans. Med. Imaging **15**(5), 620–627 (1996)
23. Droske, M., Rumpf, M.: A variational approach to nonrigid morphological image registration. SIAM J. Appl. Math. **64**(2), 668–687 (2004)
24. Evans, L.C.: Partial differential equations and Monge Kantorovich transfer. Curr. Dev. Math. **1999**, 65–126 (1997)
25. Evans, L.C.: Partial Differential Equations, vol. 19. American Mathematical Society, Providence (2002)
26. Fischer, B., Modersitzki, J.: Curvature based image registration. J. Math. Imaging Vis. **18**(1), 81–85 (2003)
27. Fischer, B., Modersitzki, J.: Ill-posed medicine – an introduction to image registration. Inverse Probl. **24**(3), 034008 (2008)
28. Fischler, M.A., Elschlager, R.A.: The representation and matching of pictorial structures. IEEE Trans. Comput. **22**, 67–92 (1973)
29. Fitzpatrick, J.M.: The existence of geometrical density-image transformations corresponding to object motion. Comput. Vis. Graph. Image Process. **44**(2), 155–174 (1988)
30. Gigengack, F., Ruthotto, L., Burger, M., Wolters, C.H., Jiang, X., Schäfers, K.P.: Motion correction in dual gated cardiac PET using mass-preserving image registration. IEEE Trans. Med. Imaging **31**(3), 698–712 (2012)
31. Glasbey, C.A., Mardia, K.V.: A review of image-warping methods. J. Appl. Stat. **25**(2), 155–171 (1998)
32. Gockenbach, M.S.: Understanding and Implementing the Finite Element Method. Society for Industrial Mathematics (SIAM), New York (2006)
33. Goshtasby, A.: 2D and 3D Image Registration. Wiley, New York (2005)
34. Haber, E., Modersitzki, J.: Numerical methods for volume preserving image registration. Inverse Probl. **20**(5), 1621–1638 (2004)
35. Haber, E., Modersitzki, J.: A scale space method for volume preserving image registration. In: Scale Space and PDE Methods in Computer Vision, pp. 561–572 (2005)
36. Haber, E., Modersitzki, J.: Image registration with guaranteed displacement regularity. Int. J. Comput. Vis. **71**(3), 361–372 (2006)
37. Haber, E., Modersitzki, J.: A multilevel method for image registration. SIAM J. Sci. Comput. **27**(5), 1594–1607 (2006)
38. Haber, E., Heldmann, S., Modersitzki, J.: A framework for image-based constrained registration with an application to local rigidity. Numer. Linear Alg. Appl. **431**(2–3), 459–470 (2007)

39. Haber, E., Heldmann, S., Modersitzki, J.: Adaptive mesh refinement for nonparametric image registration. SIAM J. Sci. Comput. **30**(6), 3012–3027 (2008)
40. Haber, E., Horesh, R., Modersitzki, J.: Numerical optimization for constrained image registration. Numer. Linear Alg. Appl. **17**(2–3), 343–359 (2010)
41. Hadamard, J.: Sur les problémes aux dérivées partielles et leur signification physique. Princeton Univ. Bull. **13**, 1–20 (1902)
42. Heldmann, S.: Non-Linear Registration Based on Mutual Information. Ph.D. thesis, Institute of Mathematics, University of Lübeck (2006)
43. Henn, S.: A multigrid method for a fourth-order diffusion equation with application to image processing. SIAM J. Sci. Comput. **27**(3), 831 (2005)
44. Hestenes, M.R., Stiefel, E.: Methods of conjugate gradients for solving linear systems. J. Res. Natl. Bureau Stand. **49**(6), 409–439 (1952)
45. Hill, D.L.G., Batchelor, P., Holden, M., Hawkes, D.J.: Medical image registration. Phys. Med. Biol. **46**, R1–R45 (2001)
46. Hinterberger, W., Scherzer, O., Schnörr, C., Weickert, J.: Analysis of optical flow models in the framework of the calculus of variations. Numer. Funct. Anal. Optim. **23**(1), 69–90 (2002)
47. Holland, D., Kuperman, J.M., Dale, A.M.: Efficient correction of inhomogeneous static magnetic field-induced distortion in echo planar imaging. NeuroImage **50**(1), 175–183 (2010)
48. Horn, B.K.P., Schunck, B.G.: Determining optical flow. Artif. Intell. **17**(1–3), 185–203 (1981)
49. Kantorovich, L.V.: On a problem of monge. Uspekhi Mat. Nauk. **3**, 225–226 (1948)
50. Keeling, S.L., Ring, W.: Medical image registration and interpolation by optical flow with maximal rigidity. J. Math. Imaging Vis. **23**(1), 47–65 (2005)
51. Lucas, B.D., Kanade, T.: An iterative image registration technique with an application to stereo vision. In: Proceedings of the 7th International Joint Conference on Artificial Intelligence (1981)
52. Maintz, J., Viergever, M.A.: A survey of medical image registration. Med. Image Anal. **2**(1), 1–36 (1998)
53. Modersitzki, J.: Numerical Methods for Image Registration. Oxford University Press, Oxford (2004)
54. Modersitzki, J.: Flirt with rigidity-image registration with a local non-rigidity penalty. Int. J. Comput. Vis. **76**(2), 153–163 (2008)
55. Modersitzki, J.: FAIR: Flexible Algorithms for Image Registration. Society for Industrial and Applied Mathematics, Philadelphia (2009)
56. Modersitzki, J., Haber, E.: Intensity gradient based registration and fusion of multi-modal images. Methods Inform. Med. **46**(3), 292–299 (2007)
57. Nocedal, J., Wright, S.: Numerical Optimization. Springer Series in Operations Research, 2nd edn. Springer, Berlin (2006)
58. Pöschl, C., Modersitzki, J., Scherzer, O.: A variational setting for volume constrained image registration. Inverse Probl. Imaging **4**(3), 505–522 (2010)
59. Rohlfing, T., Maurer, C.R., Bluemke, D.A., Jacobs, M.A.: Volume-preserving nonrigid registration of MR breast images using free-form deformation with an incompressibility constraint. IEEE Trans. Med. Imaging **22**(6), 730–741 (2003)
60. Ruthotto, L.: Mass-preserving registration of medical images. German diploma thesis (mathematics), University of Münster (2010)
61. Ruthotto, L., Gigengack, F., Burger, M., Wolters, C.H., Jiang, X., Schäfers, K.P., Modersitzki, J.: A simplified pipeline for motion correction in dual gated cardiac PET. In: Tolxdorff, T., Deserno, T.M., Handels, H., Meinzer, H.-P. (eds.) Bildverarbeitung für die Medizin 2012, pp. 51–56. Springer, Berlin (2012)
62. Ruthotto, L., Kugel, H., Olesch, J., Fischer, B., Modersitzki, J., Burger, M., Wolters, C.H.: Diffeomorphic susceptibility artefact correction of diffusion-weighted magnetic resonance images. Phys. Med. Biol. **57**(18), 5715–5731 (2012)
63. Ruthotto, L., Mohammadi, S., Heck, C., Modersitzki, J., Weiskopf, N.: HySCO – Hyperelastic Susceptibility Artifact Correction of DTI in SPM. In: Bildverarbeitung für die Medizin 2013 (2013)

64. Skare, S., Andersson, J.L.R.: Correction of MR image distortions induced by metallic objects using a 3D cubic B-spline basis set: application to stereotactic surgical planning. Magn. Reson. Med. **54**(1), 169–181 (2005)
65. Staring, M., Klein, S., Pluim, J.P.W.: A rigidity penalty term for nonrigid registration. Med. Phys. **34**(11), 4098–4108 (2007)
66. Stehling, M., Turner, R., Mansfield, P.: Echo-planar imaging: magnetic resonance imaging in a fraction of a second. Science **254**(5028), 43–50 (1991)
67. Tikhonov, A.: Solution of incorrectly formulated problems and the regularization method. Soviet Math. Dokl. **4**, 1035–1038 (1963)
68. Tikhonov, A., Arsenin, V.: Solution of Ill-Posed Problems. Winston & Sons, Washington (1977)
69. Trouvé, A.: Diffeomorphisms groups and pattern matching in image analysis. Int. J. Comput. Vis. **28**(3), 213–221 (1998)
70. Vercauteren, T., Pennec, X., Perchant, A., Ayache, N.: Diffeomorphic demons: efficient non-parametric image registration. NeuroImage **45**(1), 61–72 (2009)
71. Viola, P., Wells, W.M.: Alignment by maximization of mutual information. Int. J. Comput. Vis. **24**(2), 137–154 (1997)
72. Yanovsky, I., Le Guyader, C., Leow, A., Toga, A., Thompson, P.M., Vese, L.: Unbiased volumetric registration via nonlinear elastic regularization. In: Mathematical Foundations of Computational Anatomy (2008)
73. Zitova, B., Flusser, J.: Image registration methods: a survey. Image Vis. Comput. **21**(11), 977–1000 (2003)

Starlet Transform in Astronomical Data Processing

Jean-Luc Starck, Fionn Murtagh, and Mario Bertero

Contents

1 Introduction 2054
Source Detection 2055
2 Standard Approaches to Source Detection 2056
The Traditional Data Model 2057
PSF Estimation 2057
Background Estimation 2058
Convolution 2058
Detection 2058
Deblending/Merging 2059
Photometry and Classification 2059
3 Mathematical Modeling 2061
Sparsity Data Model 2061
The Starlet Transform 2062
The Starlet Reconstruction 2064
Starlet Transform: Second Generation 2066
Sparse Modeling of Astronomical Images 2069
Sparse Positive Decomposition 2071
4 Source Detection Using a Sparsity Model 2073
Detection Through Wavelet Denoising 2073
The Multiscale Vision Model 2075
Source Reconstruction 2079
Examples 2080

J.-L. Starck (✉)
CEA, Laboratoire AIM, CEA/DSM-CNRS-Université Paris Diderot, CEA, IRFU, Service d'Astrophysique, Centre de Saclay, Gif-Sur-Yvette Cedex, France
e-mail: jstarck@cea.fr

F. Murtagh
School of Computer Science and Informatics, De Montfort University, Leicester, UK
e-mail: fmurtagh@acm.org

M. Bertero
DIBRIS, Università di Genova, Genova, Italy
e-mail: bertero@disi.unige.it

© Springer Science+Business Media New York 2015

O. Scherzer (ed.), *Handbook of Mathematical Methods in Imaging*,
DOI 10.1007/978-1-4939-0790-8_34

5 Deconvolution ... 2081
Statistical Approach to Deconvolution ... 2083
The Richardson–Lucy Algorithm ... 2087
Blind Deconvolution ... 2089
Deconvolution with a Sparsity Prior ... 2090
Detection and Deconvolution ... 2092
Object Reconstruction Using the PSF ... 2092
The Algorithm ... 2093
Space-Variant PSF ... 2094
Undersampled Point Spread Function ... 2094
Example: Application to Abell 1689 ISOCAM Data ... 2094
6 Conclusion ... 2095
Cross-References ... 2095
Recommended Readings ... 2096
References ... 2096

Abstract

We begin with traditional source detection algorithms in astronomy. We then introduce the sparsity data model. The starlet wavelet transform serves as our main focus in this article. Sparse modeling and noise modeling are described. Applications to object detection and characterization, and to image filtering and deconvolution, are discussed. The multiscale vision model is a further development of this work, which can allow for image reconstruction when the point spread function is not known or not known well. Bayesian and other algorithms are described for image restoration. A range of examples is used to illustrate the algorithms.

1 Introduction

Data analysis is becoming more and more important in astronomy. This can be explained by detector evolution, which concerns all wavelengths. In the 1980s, CCD (charge-coupled device) images had a typical size of 512×512 pixels, while astronomers now have CCD mosaics with $16{,}000 \times 16{,}000$ pixels or even more. At the same time, methods of analysis have become much more complex, and the human and financial efforts to create and process the data can sometimes be of the same order as for the construction of the instrument itself. As an example, for the ISOCAM camera of the Infrared Space Observatory (ISO), the command software of the instrument, and the online and offline data processing, required altogether 70 person years of development, while 200 person years were necessary for the construction of the camera. The data analysis effort for the PLANCK project is even larger. Furthermore, the quantity of outputs requires the use of databases, and in parallel sophisticated tools are needed to extract ancillary astrophysical information, generally now through the web. From the current knowledge, new questions emerge, and it is necessary to proceed to new observations of a given object or a part of the sky. The acquired data need to be calibrated prior to useful information for the scientific project being extracted.

Data analysis acts during the calibration, the scientific information extraction process, and the database manipulation. The calibration phase consists of correcting various instrumental effects, such as the dark current (i.e., in the absence of any light, the camera does not return zero values, and the measured image is called the dark image and needs to be subtracted from any observation) or the flat-field correction (i.e., for uniform light, the detector does not return the same value for each pixel, and a normalization needs to be performed by dividing the observed image by the "flat" image). Hence, it is very important to know well the parameters of the detector (flat-field image, dark image, etc.), because any error on these parameters will propagate to the measurements. Other effects can also be corrected during this phase, such as the removal of the cosmic ray impacts or the field distortion (the pixel surface for each pixel does not correspond to the same surface on the sky). Depending on the knowledge of the instrument, each of these tasks may be more or less difficult.

Once the data are calibrated, the analysis phase can start. Following the scientific objectives, several kinds of information can be extracted from the data, such as the detection of stars and galaxies, the measurement of their position, intensity, and various morphological parameters. The results can be compared to existing catalogs, obtained from previous observations. It is obviously impossible to cite all operations we may want to carry through on an astronomical image, and we have just mentioned the most common. In order to extract the information, it is necessary to take into account noise and point spread function. Noise is the random fluctuation which is added to the CCD data and comes partially from the detector and partially from the data. In addition to the errors induced by the noise on the different measurements, noise also limits the detection of objects and can be responsible for false detections. The point spread function is manifested in how the image of a star, for example, is generally spread out on several pixels, caused by the atmosphere's effect on the light path. The main effect is a loss of resolution, because two sufficiently close objects cannot be separated. Once information has been extracted, such details can be compared to our existing knowledge. This comparison allows us to validate or reject our understanding of the universe.

In this chapter, we will discuss in detail how to detect objects in astronomical images and how to take into account the point spread function through the deconvolution processing.

Source Detection

As explained above, source (i.e., object) extraction from images is a fundamental step for astronomers. For example, to build catalogs, stars and galaxies must be identified and their position and photometry must be estimated with good accuracy. Catalogs comprise a key result of astronomical research. Various methods have been proposed to support the construction of catalogs. One of the now most widely used software packages is SExtractor [6], which is capable of handling very large images. A standard source detection approach, such as in SExtractor, consists of the following steps:

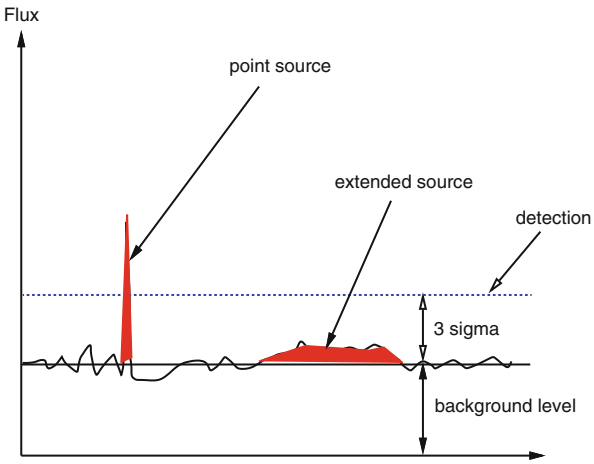


Fig. 1 Example of astronomical data: a point source and an extended source are shown, with noise and background. The extended object, which can be detected by eye, is undetected by a standard detection approach

1. Background estimation.
2. Convolution with a mask.
3. Detection.
4. Deblending/merging.
5. Photometry.
6. Classification.

These different steps are described in the next section. Astronomical images contain typically a large set of point-like sources (the stars), some quasi point-like objects (faint galaxies, double stars), and some complex and diffuse structures (galaxies, nebulae, planetary stars, clusters, etc.). These objects are often hierarchically organized: a star in a small nebula, itself embedded in a galaxy arm, itself included in a galaxy, and so on.

The standard approach, which is presented in detail in Sect. 2, presents some limits, when we are looking for faint extended objects embedded in noise. Figure 1 shows a typical example where a faint extended object is under the detection limit. In order to detect such objects, more complex data modeling needs to be defined. Section 3 presents another approach to model and represent astronomical data, by using a sparse model in a wavelet dictionary. A specific wavelet transform, called the *starlet transform* or the isotropic undecimated wavelet transform, is presented. Based on this new modeling, several approaches are proposed in Sects. 4 and 5.

2 Standard Approaches to Source Detection

We describe here the most popular way to create a catalog of galaxies from astronomical images.

The Traditional Data Model

The observed data Y can be decomposed into two parts, the signal X and the noise N:

$$Y[k,l] = X[k,l] + N[k,l] \tag{1}$$

The imaging system can also be considered. If it is linear, the relation between the data and the image in the same coordinate frame is a convolution:

$$Y[k,l] = (HX)[k,l] + N[k,l] \tag{2}$$

where H is the matrix related to the point spread function (PSF) of the imaging system.

In most cases, objects of interest are superimposed on a relatively flat signal B, called *background signal*. The model becomes

$$Y[k,l] = (HX)[k,l] + B[k,l] + N[k,l] \tag{3}$$

PSF Estimation

The PSF H can be estimated from the data or from an optical model of the imaging telescope. In astronomical images, the data may contain stars, or one can point towards a reference star in order to reconstruct a PSF. The drawback is the "degradation" of this PSF because of unavoidable noise or spurious instrument signatures in the data. So, when reconstructing a PSF from experimental data, one has to reduce very carefully the images used (background removal for instance). Another problem arises when the PSF is highly variable with time, as is the case for adaptive optics (AO) images. This means usually that the PSF estimated when observing a reference star, after or before the observation of the scientific target, has small differences from the perfectly correct PSF.

Another approach consists of constructing a synthetic PSF. Various studies [11, 21, 38, 39] have suggested a radially symmetric approximation to the PSF:

$$P(r) \propto \left(1 + \frac{r^2}{R^2}\right)^{-\beta} \tag{4}$$

The parameters β and R are obtained by fitting the model with stars contained in the data.

In the case of AO systems, this model can be used for the tail of the PSF (the so-called *seeing* contribution), while in the central region, the system provides an approximation of the diffraction-limited PSF. The quality of the approximation is measured by the Strehl ratio (SR), which is defined as the ratio of the observed peak

intensity in the image of a point source to the theoretical peak intensity of a perfect imaging system.

Background Estimation

The background must be accurately estimated; otherwise it will introduce bias in flux estimation. In [7,28], the image is partitioned into blocks, and the local sky level in each block is estimated from its histogram. The pixel intensity histogram $p(Y)$ is modeled using three parameters, the true sky level B, the RMS (root mean square) noise σ, and a parameter describing the asymmetry in $p(Y)$ due to the presence of objects, and is defined by [7]:

$$p(Y) = \frac{1}{a} \exp\left(\sigma^2/2a^2\right) \exp\left[-(Y - B)/a\right] \operatorname{erfc}\left(\frac{\sigma}{a} - \frac{(Y - B)}{\sigma}\right) \tag{5}$$

Median filtering can be applied to the 2D array of background measurements in order to correct for spurious background values. Finally the background map is obtained by a bilinear or a cubic interpolation of the 2D array. The block size is a crucial parameter. If it is too small, the background estimation map will be affected by the presence of objects, and if too large it will not take into account real background variations.

In [6, 15], the local sky level is calculated differently. A 3-sigma clipping around the median is performed in each block. If the standard deviation is changed by less than 20 % in the clipping iterations, the block is uncrowded, and the background level is considered to be equal to the mean of the clipped histogram. Otherwise, it is calculated by $c_1 \times \text{median} - c_2 \times \text{mean}$, where $c_1 = 3, c_2 = 2$ in [15] and $c_1 = 2.5, c_2 = 1.5$ in [6]. This approach has been preferred to histogram fitting for two reasons: it is more efficient from the computation point of view and more robust with small sample size.

Convolution

In order to optimize the detection, the image must be convolved with a filter. The shape of this filter optimizes the detection of objects with the same shape. Therefore, for star detection, the optimal filter is the PSF. For extended objects, a larger filter size is recommended. In order to have optimal detection for any object size, the detection must be repeated several times with different filter sizes, leading to a kind of multiscale approach.

Detection

Once the image is convolved, all pixels $Y[k, l]$ at location (k, l) with a value larger than $T[k, l]$ are considered as significant, i.e., belonging to an object. $T[k, l]$ is

generally chosen as $B[k,l] + K\sigma$, where $B[k,l]$ is the background estimate at the same position, σ is the noise standard deviation, and K is a given constant (typically chosen between 3 and 5). The thresholded image is then segmented, i.e., a label is assigned to each group of connected pixels. The next step is to separate the blended objects which are connected and have the same label.

An alternative to the thresholding/segmentation procedure is to find peaks. This is only well suited to star detection and not to extended objects. In this case, the next step is to merge the pixels belonging to the same object.

Deblending/Merging

This is the most delicate step. Extended objects must be considered as single objects, while multiple objects must be well separated. In SExtractor, each group of connected pixels is analyzed at different intensity levels, starting from the highest down to the lowest level. The pixel group can be seen as a surface, with mountains and valleys. At the beginning (highest level), only the highest peak is visible. When the level decreases, several other peaks may become visible, defining therefore several structures. At a given level, two structures may become connected, and the decision whether they form only one (i.e., merging) or several objects (i.e., deblending) must be taken. This is done by comparing the integrated intensities inside the peaks. If the ratio between them is too low, then the two structures must be merged.

Photometry and Classification

Photometry

Several methods can be used to derive the photometry of a detected object [7, 29]. Adaptive aperture photometry uses the first image moment to determine the elliptical aperture from which the object flux is integrated. Kron [29] proposed an aperture size of twice the radius of the first image moment radius r_1, which leads to recovery of most of the flux (>90 %). In [6], the value of $2.5r_1$ is discussed, leading to loss of less than 6 % of the total flux. Assuming that the intensity profiles of the faint objects are Gaussian, flux estimates can be refined [6, 35]. When the image contains only stars, specific methods can be developed which take the PSF into account [18, 42].

Star–Galaxy Separation

In the case of star–galaxy classification, following the scanning of digitized images, Kurtz [30] lists the following parameters which have been used:

1. Mean surface brightness;
2. Maximum intensity and area;
3. Maximum intensity and intensity gradient;

4. Normalized density gradient;
5. Areal profile;
6. Radial profile;
7. Maximum intensity, 2nd and 4th order moments, and ellipticity;
8. The fit of galaxy and star models;
9. Contrast versus smoothness ratio;
10. The fit of a Gaussian model;
11. Moment invariants;
12. Standard deviation of brightness;
13. 2nd order moment;
14. Inverse effective squared radius;
15. Maximum intensity and intensity-weighted radius;
16. 2nd and 3rd order moments, number of local maxima, and maximum intensity.

References for all of these may be found in the cited work. Clearly there is room for differing views on parameters to be chosen for what is essentially the same problem. It is of course the case also that aspects such as the following will help to orientate us towards a particular set of parameters in a particular case: the quality of the data; the computational ease of measuring certain parameters; the relevance and importance of the parameters measured relative to the data analysis output (e.g., the classification, or the planar graphics); and, similarly, the importance of the parameters relative to theoretical models under investigation.

Galaxy Morphology Classification

The inherent difficulty of characterizing spiral galaxies especially when not face-on has meant that most work focuses on ellipticity in the galaxies under study. This points to an inherent bias in the potential multivariate statistical procedures. In the following, it will not be attempted to address problems of galaxy photometry per se [17,44], but rather to draw some conclusions on what types of parameters or features have been used in practice.

From the point of view of multivariate statistical algorithms, a reasonably homogeneous set of parameters is required. Given this fact, and the available literature on quantitative galaxy morphological classification, two approaches to parameter selection appear to be strongly represented:

1. The luminosity profile along the major axis of the object is determined at discrete intervals. This may be done by the fitting of elliptical contours, followed by the integrating of light in elliptical annuli [33]. A similar approach was used in the ESO–Uppsala survey. Noisiness and faintness require attention to robustness in measurement: the radial profile may be determined taking into account the assumption of a face-on optically thin axisymmetric galaxy and may be further adjusted to yield values for circles of given radius [64]. Alternatively, isophotal contours may determine the discrete radial values for which the profile is determined [62].

2. Specific morphology-related parameters may be derived instead of the profile. The integrated magnitude within the limiting surface brightness of 25 or 26 mag. arcsec^{-2} in the visual is popular [33, 61]. The logarithmic diameter (D_{26}) is also supported by Okamura [43]. It may be interesting to fit to galaxies under consideration model bulges and disks using, respectively, $r^{\frac{1}{4}}$ or exponential laws [62], in order to define further parameters. Some catering for the asymmetry of spirals may be carried out by decomposing the object into octants; furthermore, the taking of a Fourier transform of the intensity may indicate aspects of the spiral structure [61].

The following remarks can be made relating to image data and reduced data:

- The range of parameters to be used should be linked to the subsequent use to which they might be put, such as to underlying physical aspects.
- Parameters can be derived from a carefully constructed luminosity profile, rather than it being possible to derive a profile from any given set of parameters.
- The presence of both partially reduced data such as luminosity profiles, and more fully reduced features such as integrated flux in a range of octants, is of course not a hindrance to analysis. However, it is more useful if the analysis is carried out on both types of data separately.

Parameter data can be analyzed by clustering algorithms, by principal component analysis, or by methods for discriminant analysis. Profile data can be sampled at suitable intervals and thus analyzed also by the foregoing procedures. It may be more convenient in practice to create dissimilarities between profiles and analyze these dissimilarities: this can be done using clustering algorithms with dissimilarity input.

3 Mathematical Modeling

Different models may be considered to represent the data. One of the most effective is certainly the sparsity model, especially when a specific wavelet dictionary is chosen to represent the data. We introduce here the sparsity concept as well as the wavelet transform decomposition, which is the most used in astronomy.

Sparsity Data Model

A signal X, $X = [x_1, \ldots, x_N]^T$, is sparse if most of its entries are equal to zero. For instance, a k-sparse signal is a signal where only k samples have a nonzero value. A less strict definition is to consider a signal as weakly sparse or compressible when only a few of its entries have a large magnitude, while most of them are close to zero.

If a signal is not sparse, it may be *sparsified* using a given data representation. For instance, if X is a sine, it is clearly not sparse but its Fourier transform is extremely sparse (i.e., 1-sparse). Hence, we say that a signal X is sparse in the Fourier domain if its Fourier coefficients $\hat{X}[u]$, $\hat{X}[u] = \frac{1}{N}\sum_{k=-\infty}^{+\infty} X[k]e^{2i\pi\frac{uk}{N}}$, are sparse. More generally, we can model a vector signal $X \in \mathbb{R}^N$ as the linear combination of T elementary waveforms, also called *signal atoms*: $X = \mathbf{\Phi}\alpha = \sum_{i=1}^{T}\alpha[i]\phi_i$, where $\alpha[i] = \langle X, \phi_i\rangle$ are called the decomposition coefficients of X in the dictionary $\mathbf{\Phi} = [\phi_1, \ldots, \phi_T]$ (the $N \times T$ matrix whose columns are the atoms normalized to a unit ℓ_2-norm, i.e., $\forall i \in [1, T], \|\phi_i\|_{\ell^2} = 1$).

Therefore, to get a sparse representation of our data, we need first to define the dictionary $\mathbf{\Phi}$ and then to compute the coefficients α. x is sparse in Φ if the sorted coefficients in decreasing magnitude have fast decay, i.e., most coefficients α vanish except for a few.

The best dictionary is the one which leads to the sparsest representation. Hence, we could imagine having a huge overcomplete dictionary (i.e., $T \gg N$), but we would be faced with prohibitive computation time cost for calculating the α coefficients. Therefore, there is a trade-off between the complexity of our analysis step (i.e., the size of the dictionary) and the computation time. Some specific dictionaries have the advantage of having fast operators and are very good candidates for analyzing the data.

The Isotropic Undecimated Wavelet Transform (IUWT), also called *starlet wavelet transform*, is well known in the astronomical domain because it is well adapted to astronomical data where objects are more or less isotropic in most cases [54, 57]. For most astronomical images, the starlet dictionary is very well adapted.

The Starlet Transform

The starlet wavelet transform [53] decomposes an $n \times n$ image c_0 into a coefficient set $W = \{w_1, \ldots, w_J, c_J\}$, as a superposition of the form

$$c_0[k, l] = c_J[k, l] + \sum_{j=1}^{J} w_j[k, l],$$

where c_J is a coarse or smooth version of the original image c_0 and w_j represents the details of c_0 at scale 2^{-j} (see Starck et al. [56, 58] for more information). Thus, the algorithm outputs $J + 1$ sub-band arrays of size $N \times N$. (The present indexing is such that $j = 1$ corresponds to the finest scale or high frequencies.)

The decomposition is achieved using the filter bank $(h_{2D}, g_{2D} = \delta - h_{2D}, \tilde{h}_{2D} = \delta, \tilde{g}_{2D} = \delta)$, where h_{2D} is the tensor product of two 1D filters h_{1D} and δ is the Dirac function. The passage from one resolution to the next one is obtained using the "à trous" ("with holes") algorithm [58]:

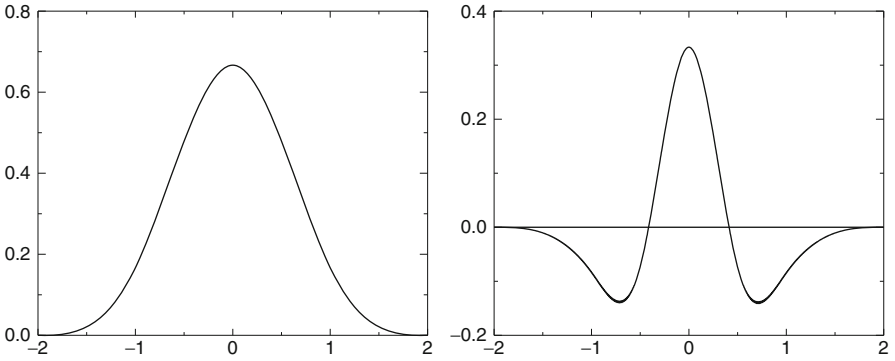


Fig. 2 *Left*, the cubic spline function ϕ; *right*, the wavelet ψ

$$
\begin{aligned}
c_{j+1}[k,l] &= \sum_m \sum_n h_{1\mathrm{D}}[m]\, h_{1\mathrm{D}}[n]\, c_j[k+2^j m, l+2^j n], \\
w_{j+1}[k,l] &= c_j[k,l] - c_{j+1}[k,l]\ ,
\end{aligned}
\tag{6}
$$

If we choose a B_3-spline for the scaling function,

$$
\phi(x) = B_3(x) =
$$

$$
\frac{1}{12}(|\, x-2\,|^3 -4\,|\, x-1\,|^3 +6\,|\, x\,|^3 -4\,|\, x+1\,|^3 +|\, x+2\,|^3) \tag{7}
$$

the coefficients of the convolution mask in one dimension are $h_{1\mathrm{D}} = \{\frac{1}{16}, \frac{1}{4}, \frac{3}{8}, \frac{1}{4}, \frac{1}{16}\}$, and in two dimensions,

$$
h_{2\mathrm{D}} = \begin{pmatrix} 1/16 & 1/4 & 3/8 & 1/4 & 1/16 \end{pmatrix}
\begin{pmatrix} \frac{1}{16} \\ \frac{1}{4} \\ \frac{3}{8} \\ \frac{1}{4} \\ \frac{1}{16} \end{pmatrix} =
\begin{pmatrix}
\frac{1}{256} & \frac{1}{64} & \frac{3}{128} & \frac{1}{64} & \frac{1}{256} \\
\frac{1}{64} & \frac{1}{16} & \frac{3}{32} & \frac{1}{16} & \frac{1}{64} \\
\frac{3}{128} & \frac{3}{32} & \frac{9}{64} & \frac{3}{32} & \frac{3}{128} \\
\frac{1}{64} & \frac{1}{16} & \frac{3}{32} & \frac{1}{16} & \frac{1}{64} \\
\frac{1}{256} & \frac{1}{64} & \frac{3}{128} & \frac{1}{64} & \frac{1}{256}
\end{pmatrix}
$$

Figure 2 shows the scaling function and the wavelet function when a cubic spline function is chosen as the scaling function ϕ.

The most general way to handle the boundaries is to consider that $c[k+N] = c[N-k]$ ("mirror"). But other methods can be used such as periodicity ($c[k+N] = c[N]$) or continuity ($c[k+N] = c[k]$).

The starlet transform algorithm is:

1. We initialize j to 0 and we start with the data $c_j[k,l]$.
2. We carry out a discrete convolution of the data $c_j[k,l]$ using the filter ($h_{2\mathrm{D}}$), using the separability in the two-dimensional case. In the case of the B_3-

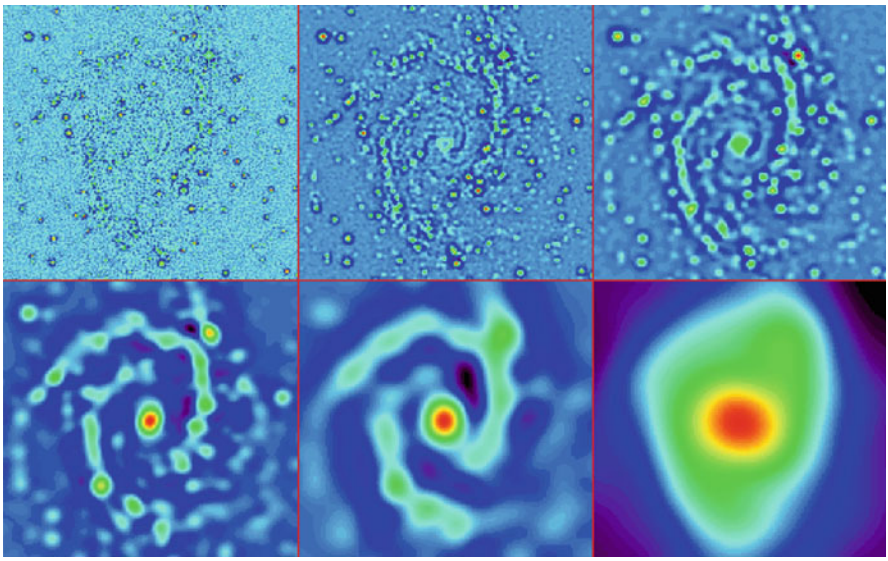


Fig. 3 Wavelet transform of NGC 2997 by the IUWT. The co-addition of these six images reproduces exactly the original image

spline, this leads to a row-by-row convolution with $\left(\frac{1}{16}, \frac{1}{4}, \frac{3}{8}, \frac{1}{4}, \frac{1}{16}\right)$, followed by column-by-column convolution. The distance between the central pixel and the adjacent ones is 2^j.

3. After this smoothing, we obtain the discrete wavelet transform from the difference $c_j[k,l] - c_{j+1}[k,l]$.
4. If j is less than the number J of resolutions we want to compute, we increment j and then go to step 2.
5. The set $\alpha = \{w_1, \ldots, w_J, c_J\}$ represents the wavelet transform of the data.

This starlet transform is very well adapted to the detection of isotropic features, and this explains its success for astronomical image processing, where the data contain mostly isotropic or quasi-isotropic objects, such as stars, galaxies, or galaxy clusters.

Figure 3 shows the starlet transform of the galaxy NGC 2997 displayed in Fig. 4. Five wavelet scales and the final smoothed plane (lower right) are shown. The original image is given exactly by the sum of these six images.

The Starlet Reconstruction

The reconstruction is straightforward. A simple co-addition of all wavelet scales reproduces the original map: $c_0[k,l] = c_J[k,l] + \sum_{j=1}^{J} w_j[k,l]$. But because the transform is non-subsampled, there are many ways to reconstruct the original image

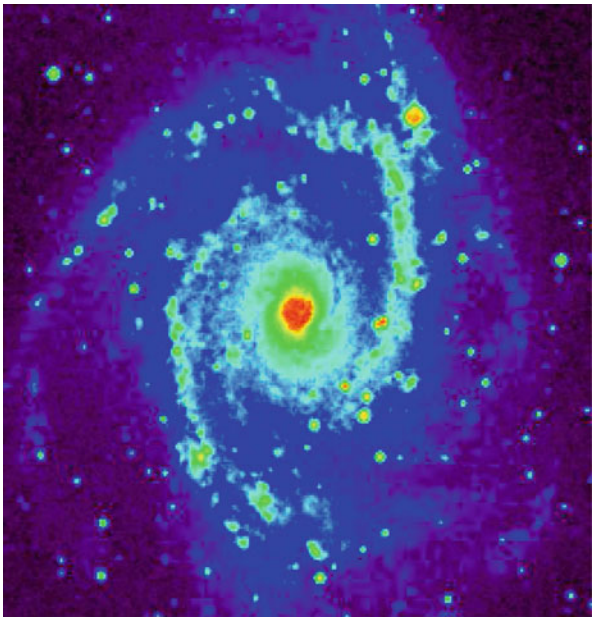


Fig. 4 Galaxy NGC 2997

from its wavelet transform [53]. For a given wavelet filter bank $(h,\ g)$, associated with a scaling function ϕ and a wavelet function ψ, any synthesis filter bank $(\tilde{h},\ \tilde{g})$, which satisfies the following reconstruction condition

$$\hat{h}^*(\nu)\hat{\tilde{h}}(\nu) + \hat{g}^*(\nu)\hat{\tilde{g}}(\nu) = 1\ , \tag{8}$$

leads to exact reconstruction. For instance, for isotropic h, if we choose $\tilde{h} = h$ (the synthesis scaling function $\tilde{\phi} = \phi$), we obtain a filter $\tilde{g}$ defined by [53]

$$\tilde{g} = \delta + h\ .$$

If h is a positive filter, then g is also positive. For instance, if $h_{1\mathrm{D}} = [1, 4, 6, 4, 1]/16$, then $\tilde{g}_{1\mathrm{D}} = [1, 4, 22, 4, 1]/16$. That is, $\tilde{g}_{1\mathrm{D}}$ *is positive*. This means that $\tilde{g}$ is no longer related to a wavelet function. The 1D detail synthesis function related to $\tilde{g}_{1\mathrm{D}}$ is defined by

$$\frac{1}{2}\tilde{\psi}_{1\mathrm{D}}\left(\frac{t}{2}\right) = \phi_{1\mathrm{D}}(t) + \frac{1}{2}\phi_{1\mathrm{D}}\left(\frac{t}{2}\right)\ . \tag{9}$$

Note that by choosing $\tilde{\phi}_{1\mathrm{D}} = \phi_{1\mathrm{D}}$, any synthesis function $\tilde{\psi}_{1\mathrm{D}}$ which satisfies

$$\hat{\tilde{\psi}}_{1\mathrm{D}}(2\nu)\hat{\psi}_{1\mathrm{D}}(2\nu) = \hat{\phi}^2_{1\mathrm{D}}(\nu) - \hat{\phi}^2_{1\mathrm{D}}(2\nu) \tag{10}$$

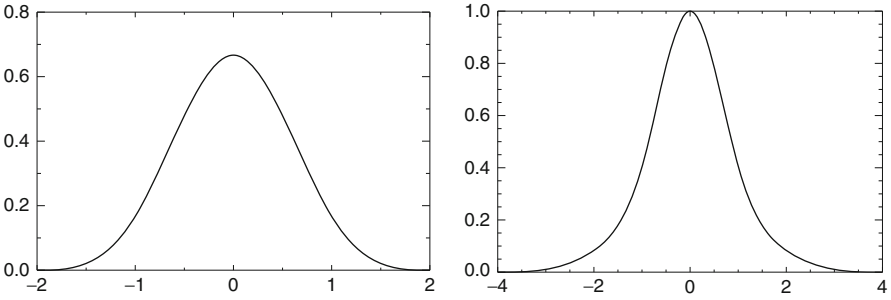


Fig. 5 *Left*, $\tilde{\phi}_{1D}$ the 1D synthesis scaling function and *right*, $\tilde{\psi}_{1D}$ the 1D detail synthesis function

leads to an exact reconstruction [36] and $\hat{\tilde{\psi}}_{1D}(0)$ can take any value. The synthesis function $\tilde{\psi}_{1D}$ does not need to verify the admissibility condition (i.e., to have a zero mean).

Figure 5 shows the two functions $\tilde{\phi}_{1D}(= \phi_{1D})$ and $\tilde{\psi}_{1D}$ used in the reconstruction in 1D, corresponding to the synthesis filters $\tilde{h}_{1D} = h_{1D}$ and $\tilde{g}_{1D} = \delta + h_{1D}$. More details can be found in [53].

Starlet Transform: Second Generation

A particular case is obtained when $\hat{\tilde{\phi}}_{1D} = \hat{\phi}_{1D}$ and $\hat{\psi}_{1D}(2\nu) = \frac{\hat{\phi}^2_{1D}(\nu) - \hat{\phi}^2_{1D}(2\nu)}{\hat{\phi}_{1D}(\nu)}$, which leads to a filter g_{1D} equal to $\delta - h_{1D} \star h_{1D}$. In this case, the synthesis function $\tilde{\psi}_{1D}$ is defined by $\frac{1}{2}\tilde{\psi}_{1D}(\frac{t}{2}) = \phi_{1D}(t)$, and the filter $\tilde{g}_{1D} = \delta$ is the solution to (8).

We end up with a synthesis scheme where only the smooth part is convolved during the reconstruction.

Deriving h from a spline scaling function, for instance $B_1(h_1 = [1, 2, 1]/4)$ or B_3 $(h_3 = [1, 4, 6, 4, 1]/16)$ (note that $h_3 = h_1 \star h_1$), since h_{1D} is even-symmetric (i.e., $H(z) = H(z^{-1})$), the z-transform of g_{1D} is then

$$\begin{aligned} G(z) &= 1 - H^2(z) = 1 - z^4 \left(\frac{1 + z^{-1}}{2} \right)^8 \\ &= \frac{1}{256} \left(-z^4 - 8z^3 - 28z^2 - 56z + 186 - 56z^{-1} - 28z^{-2} - 8z^{-3} - z^{-4} \right), \end{aligned} \tag{11}$$

which is the z-transform of the filter

$$g_{1D} = [-1, -8, -28, -56, 186, -56, -28, -8, -1]/256.$$

We get the following filter bank:

$$\begin{aligned} h_{1D} &= h_3 = \tilde{h} = [1,4,6,4,1]/16 \\ g_{1D} &= \delta - h \star h = [-1,-8,-28,-56,186,-56,-28,-8,-1]/256 \\ \tilde{g}_{1D} &= \delta \,. \end{aligned} \tag{12}$$

The second-generation starlet transform algorithm is:

1. We initialize j to 0 and we start with the data $c_j[k]$.
2. We carry out a discrete convolution of the data $c_j[k]$ using the filter h_{1D}. The distance between the central pixel and the adjacent ones is 2^j. We obtain $c_{j+1}[k]$.
3. We do exactly the same convolution on $c_{j+1}[k]$ and we obtain $c'_{j+1}[k]$.
4. After this two-step smoothing, we obtain the discrete starlet wavelet transform from the difference $w_{j+1}[k] = c_j[k] - c'_{j+1}[k]$.
5. If j is less than the number J of resolutions we want to compute, we increment j and then go to step 2.
6. The set $\alpha = \{w_1, \ldots, w_J, c_J\}$ represents the starlet wavelet transform of the data.

As in the standard starlet transform, extension to 2D is trivial. We just replace the convolution with h_{1D} by a convolution with the filter h_{2D}, which is performed efficiently by using the separability.

With this specific filter bank, there is a no convolution with the filter $\tilde{g}_{1D}$ during the reconstruction. Only the low-pass synthesis filter $\tilde{h}_{1D}$ is used.

The reconstruction formula is

$$c_j[l] = (h_{1D}^{(j)} \star c_{j+1})[l] + w_{j+1}[l] \,, \tag{13}$$

and denoting $L^j = h^{(0)} \star \cdots \star h^{(j-1)}$ and $L^0 = \delta$, we have

$$c_0[l] = \left(L^J \star c_J\right)[l] + \sum_{j=1}^{J} \left(L^{j-1} \star w_j\right)[l] \,. \tag{14}$$

Each wavelet scale is convolved with a low-pass filter.

The second-generation starlet reconstruction algorithm is:

1. The set $\alpha = \{w_1, \ldots, w_J, c_J\}$ represents the input starlet wavelet transform of the data.
2. We initialize j to $J-1$ and we start with the coefficients $c_j[k]$.
3. We carry out a discrete convolution of the data $c_{j+1}[k]$ using the filter (h_{1D}). The distance between the central pixel and the adjacent ones is 2^j. We obtain $c'_{j+1}[k]$.
4. Compute $c_j[k] = c'_{j+1}[k] + w_{j+1}[k]$.

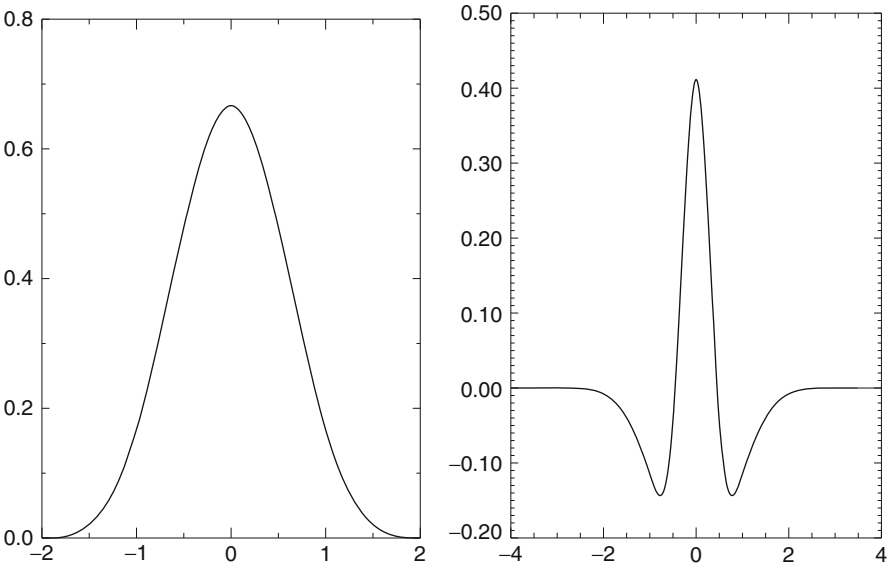


Fig. 6 *Left*, the ϕ_{1D} analysis scaling function and *right*, the ψ_{1D} analysis wavelet function. The synthesis functions $\tilde{\phi}_{1D}$ and $\tilde{\psi}_{1D}$ are the same as those in Fig. 5

5. If j is larger than 0, $j = j - 1$ and then go to step 3.
6. c_0 contains the reconstructed data.

As for the transformation, the 2D extension consists just in replacing the convolution by h_{1D} with a convolution by h_{2D}.

Figure 6 shows the analysis scaling and wavelet functions. The synthesis functions $\tilde{\phi}_{1D}$ and $\tilde{\psi}_{1D}$ are the same as those in Fig. 5. As both are positive, we have a decomposition of an image X on positive scaling functions $\tilde{\phi}_{1D}$ and $\tilde{\psi}_{1D}$, but the coefficients α are obtained with the starlet wavelet transform and have a zero mean (except for c_J), as a regular wavelet transform.

In 2D, similarly, the second-generation starlet transform leads to the representation of an image $X[k,l]$:

$$X[k,l] = \sum_{m,n} \phi^{(1)}_{j,k,l}(m,n) c_J[m,n] + \sum_{j=1}^{J} \sum_{m,n} \phi^{(2)}_{j,k,l}(m,n) w_j[m,n] \,, \qquad (15)$$

where $\phi^{(1)}_{j,k,l}(m,n) = 2^{-2j} \tilde{\phi}_{1D}(2^{-j}(k-m)) \tilde{\phi}_{1D}(2^{-j}(l-n))$ and $\phi^{(2)}_{j,k,l}(m,n) = 2^{-2j} \tilde{\psi}_{1D}(2^{-j}(k-m)) \tilde{\psi}_{1D}(2^{-j}(l-n))$.

$\phi^{(1)}$ and $\phi^{(2)}$ are positive, and w_j are zero mean 2D wavelet coefficients.

The advantage of the second-generation starlet transform will be seen in section "Sparse Positive Decomposition" below.

Sparse Modeling of Astronomical Images

Using the sparse modeling, we now consider that the observed signal X can be considered as a linear combination of a few atoms of the wavelet dictionary $\mathbf{\Phi} = [\phi_1, \ldots, \phi_T]$. The model of Eq. 3 is then replaced by the following:

$$Y = H\mathbf{\Phi}\alpha + N + B \tag{16}$$

and $X = \mathbf{\Phi}\alpha$, and $\alpha = \{w_1, \ldots, w_J, c_J\}$. Furthermore, most of the coefficients α will be equal to zero. Positions and scales of active coefficients are unknown, but they can be estimated directly from the data Y. We define the multiresolution support M of an image Y by

$$M_j[k,l] = \begin{cases} 1 & \text{if } w_j[k,l] \text{ is significant} \\ 0 & \text{if } w_j[k,l] \text{ is not significant} \end{cases} \tag{17}$$

where $w_j[k,l]$ is the wavelet coefficient of Y at scale j and at position (k,l). Hence, M describes the set of active atoms in Y. If H is compact and not too extended, then M describes also well the active set of X. This is true because the background B is generally very smooth, and therefore, a wavelet coefficient $w_j[k,l]$ of Y, which does not belong to the coarsest scale, is only dependent on X and N (the term $< \phi_i, B >$ being equal to zero).

Selection of Significant Coefficients Through Noise Modeling

We need now to determine when a wavelet coefficient is significant. Wavelet coefficients of Y are corrupted by noise, which follows in many cases a Gaussian distribution, a Poisson distribution, or a combination of both. It is important to detect the wavelet coefficients which are "significant," i.e., the wavelet coefficients which have an absolute value too large to be due to noise.

For Gaussian noise, it is easy to derive an estimation of the noise standard deviation σ_j at scale j from the noise standard deviation, which can be evaluated with good accuracy in an automated way [55]. To detect the significant wavelet coefficients, it suffices to compare the wavelet coefficients $w_j[k,l]$ to a threshold level t_j. t_j is generally taken equal to $K\sigma_j$, and K, as noted in Sect. 2, is chosen between 3 and 5. The value of 3 corresponds to a probability of false detection of 0.27 %. If $w_j[k,l]$ is small, then it is not significant and could be due to noise. If $w_j[k,l]$ is large, it is significant:

$$\begin{array}{l} \text{if } \mid w_j[k,l] \mid \geq t_j \quad \text{then } w_j[k,l] \text{ is significant} \\ \text{if } \mid w_j[k,l] \mid < t_j \quad \text{then } w_j[k,l] \text{ is not significant} \end{array} \tag{18}$$

When the noise is not Gaussian, other strategies may be used:

- **Poisson noise:** if the noise in the data Y is Poisson, the transformation [1] $\mathcal{A}(Y) = 2\sqrt{Y + \frac{3}{8}}$ acts as if the data arose from a Gaussian white noise model, with $\sigma = 1$, under the assumption that the mean value of Y is sufficiently large. However, this transform has some limits, and it has been shown that it cannot be applied for data with less than 20 counts (due to photons) per pixel. So for X-ray or gamma ray data, other solutions have to be chosen, which manage the case of a reduced number of events or photons under assumptions of Poisson statistics.
- **Gaussian + Poisson noise:** the generalization of variance stabilization [40] is

$$\mathcal{G}(Y[k,l]) = \frac{2}{\alpha}\sqrt{\alpha Y[k,l] + \frac{3}{8}\alpha^2 + \sigma^2 - \alpha g}$$

  where α is the gain of the detector and g and σ are the mean and the standard deviation of the readout noise.
- **Poisson noise with few events using the MS-VST:** for images with very few photons, one solution consists in using the Multi-Scale Variance Stabilization Transform (MS-VST) [66]. The MS-VST combines both the Anscombe transform and the starlet transform in order to produce *stabilized* wavelet coefficients, i.e., coefficients corrupted by a Gaussian noise with a standard deviation equal to 1. In this framework, wavelet coefficients are now calculated by

$$\begin{array}{c}\text{Starlet}\\ +\\ \text{MS-VST}\end{array} \begin{cases} c_j &= \sum_m \sum_n h_{1D}[m] h_{1D}[n] \\ & \quad c_{j-1}[k + 2^{j-1}m, l + 2^{j-1}n] \\ w_j &= \mathcal{A}_{j-1}(c_{j-1}) - \mathcal{A}_j(c_j) \end{cases} \tag{19}$$

  where $\mathcal{A}_j$ is the VST operator at scale j defined by

$$\mathcal{A}_j(c_j) = b^{(j)}\sqrt{|c_j + e^{(j)}|} \tag{20}$$

  where the variance stabilization constants $b^{(j)}$ and $e^{(j)}$ only depend on the filter h_{1D} and the scale level j. They can all be precomputed once for any given h_{1D} [66]. The multiresolution support is computed from the MS-VST coefficients, considering a Gaussian noise with a standard deviation equal to 1. This stabilization procedure is also invertible as we have

$$c_0 = \mathcal{A}_0^{-1}\left[\mathcal{A}_J(a_J) + \sum_{j=1}^{J} w_j\right] \tag{21}$$

For other kinds of noise (correlated noise, nonstationary noise, etc.), other solutions have been proposed to derive the multiresolution support [57].

Sparse Positive Decomposition

Many astronomical images can be modeled as a sum of positive features, like stars and galaxies, which are more or less isotropic. The previous representation, based on the starlet transform, is well adapted to the representation of isotropic objects, but does not introduce any prior relative to the positivity of the features contained in our image. A positive and sparse modeling of astronomical images is similar to Eq. 16:

$$Y = H\mathbf{\Phi}\alpha + N + B \tag{22}$$

or

$$Y = \mathbf{\Phi}\alpha + N + B \tag{23}$$

if we do not take into account the point spread function. All coefficients in α are now positive, and all atoms in the dictionary $\mathbf{\Phi}$ are positive functions. Such a decomposition normally requires computationally intensive algorithms such as matching pursuit [37]. The second-generation starlet transform offers us a new way to perform such a decomposition. Indeed, we have seen in section "Starlet Transform: Second Generation" that, using a specific filter bank, we can decompose an image Y on a positive dictionary $\mathbf{\Phi}$ (see Fig. 5) and obtain a set of coefficients $\alpha^{(Y)}$, where $\alpha^{(Y)} = \mathbf{W}Y = \{w_1, \ldots, w_J, c_J\}$, $\mathbf{W}$ being the starlet wavelet transform operator. α coefficients are positive and negative and are obtained using the standard starlet wavelet transform algorithm. Hence, by thresholding all negative (respectively, positive) coefficients, the reconstruction is always positive (respectively, negative), since $\mathbf{\Phi}$ contains only positive atoms.

Hence, we would like to have a sparse set of positive coefficients $\tilde{\alpha}$ which verify $\mathbf{\Phi}\tilde{\alpha} = Y$. But in order to take into account the background and the noise, we need to define the constraint in the wavelet space (i.e., $\mathbf{W}\mathbf{\Phi}\tilde{\alpha} = \mathbf{W}Y = \alpha^{(Y)}$), and this constraint must be applied only to the subset of coefficients in $\alpha^{(Y)}$ which are larger than the detection level. Therefore, to get a sparse positive decomposition on $\mathbf{\Phi}$, we need to minimize

$$\tilde{\alpha} = \min_{\alpha} \| \alpha \|_1 \quad \text{s.t. } M\mathbf{W}\mathbf{\Phi}\alpha = M\alpha^{(Y)} \ , \tag{24}$$

where M is the multiresolution support defined in the previous section (i.e., $M_j[k,l] = 1$ if a significant coefficient is detected at scale j and at position (k,l), and zero otherwise). To remove the background, we have to set $M_{J+1}[k,l] = 0$ for all (k,l).

It was shown that such optimization problems can be efficiently solved through an iterative soft thresholding (IST) algorithm [14, 24, 52]. The following algorithm, based on the IST, allows to take into account the noise modeling through the multiresolution support and force the coefficients to be all positive:

1. Taking the second-generation starlet wavelet transform of the data Y, we obtain $\alpha^{(Y)}$.
2. From a given noise model, determine the multiresolution support M.
3. Set the number of iterations N_{iter}, the first threshold, $\lambda^{(0)} = MAX(\alpha^{(Y)})$, and the solution $\tilde{\alpha}^{(0)} = 0$.
4. For $0 = 1, N_{\text{iter}}$ do:

   - Reconstruct the image $\tilde{Y}^{(i)}$ from $\tilde{\alpha}^{(i)}$: $\tilde{Y}^{(i)} = \mathbf{\Phi}\tilde{\alpha}^{(i)}$.
   - Taking the second-generation starlet wavelet transform of the data $\tilde{Y}^{(i)}$, we obtain $\alpha^{\tilde{Y}^{(i)}} = \mathbf{W}\mathbf{\Phi}\tilde{\alpha}^{(i)}$.
   - Compute the significant residual $r^{(i)}$:

$$r^{(i)} = M\left(\alpha^{(Y)} - \alpha^{\tilde{Y}^{(i)}}\right) = M\left(\alpha^{(Y)} - \mathbf{W}\mathbf{\Phi}\tilde{\alpha}^{(i)}\right) \tag{25}$$

   - Calculate the value $\lambda^{(i)} = \lambda^{(0)}(1 - i/N_{\text{iter}})$
   - Update the solution, by adding the residual, applying a soft thresholding on positive coefficients using the threshold level $\lambda^{(i)}$, and setting all negative coefficients to zero.

$$\begin{aligned}\tilde{\alpha}^{(i+1)} &= \left(\tilde{\alpha}^{(i)} + r^{(i)} - \lambda^{(i)}\right)_+ \\ &= \left(\tilde{\alpha}^{(i)} + M\left(\alpha^{(Y)} - \mathbf{W}\mathbf{\Phi}\tilde{\alpha}^{(i)}\right) - \lambda^{(i)}\right)_+\end{aligned} \tag{26}$$

   - $i = i + 1$.

5. The set $\tilde{\alpha} = \tilde{\alpha}^{(N_{\text{iter}})}$ represents the sparse positive decomposition of the data.

The threshold parameter $\lambda^{(i)}$ decreases with the iteration number, and it plays a role similar to the cooling parameter of the simulated annealing techniques, i.e., it allows the solution to escape from local minima.

Example 1: Sparse Positive Decomposition of NGC 2997

Figure 7 shows the positive starlet decomposition, using 100 iterations, and can be compared to Fig. 3.

Example 2: Sparse Positive Starlet Decomposition of a Simulated Image

The next example compares the standard starlet transform to the positive starlet decomposition (PSD) on a simulated image.

Figure 8 shows respectively from top to bottom and left to right (a) the original simulated image, (b) the noisy data, (c) the reconstruction from the PSD coefficients, and (d) the residual between the noisy data and the PSD reconstructed image (i.e., image b–image c). Hence, the PSD reconstructed image gives a very good approximation of the original image. No structures can be seen in the residual, and all sources are well detected.

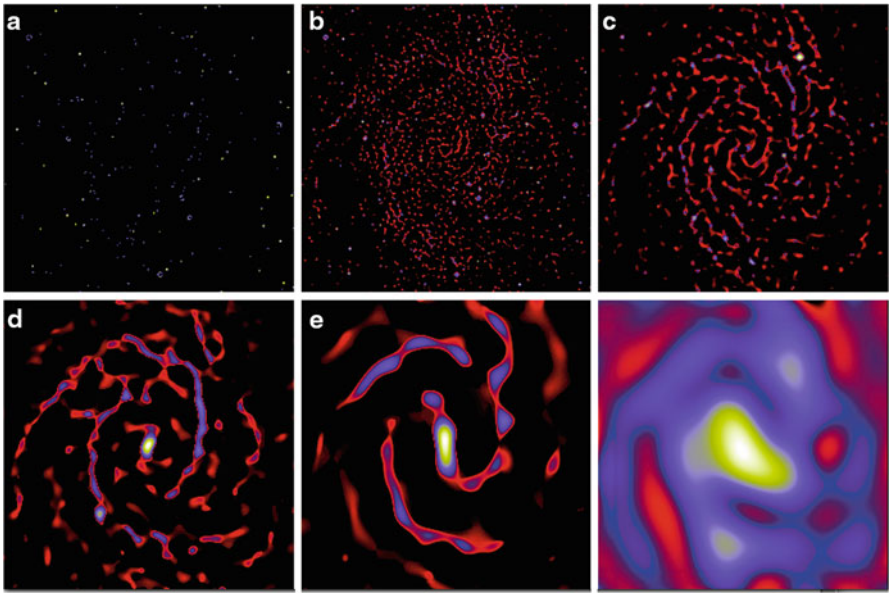


Fig. 7 Positive starlet decomposition of the galaxy NGC 2997 with six scales

The first PSD scale does not contain any nonzero coefficient. Figure 9, top, shows the first four scales of the wavelet transform, and Fig. 9, bottom, the first four scales of the PSD.

4 Source Detection Using a Sparsity Model

As described is the previous section, the wavelet coefficients of Y which do not belong to the coarsest scale c_J are not dependent on the background. This is a serious disadvantage, since the background estimation can be sometimes very problematic.

Two approaches have been proposed to detect sources, assuming the signal is sparse in the wavelet domain. The first consists in first removing the noise and the background and then applying the standard approach described in Sect. 2. It has been used for many years for X-ray source detection [45,59]. The second approach, called Multiscale Vision Model [8], attempts to define directly an astronomical object in the wavelet space.

Detection Through Wavelet Denoising

The most commonly used filtering method is hard thresholding, which consists of setting to 0 all wavelet coefficients of Y which have an absolute value lower than a

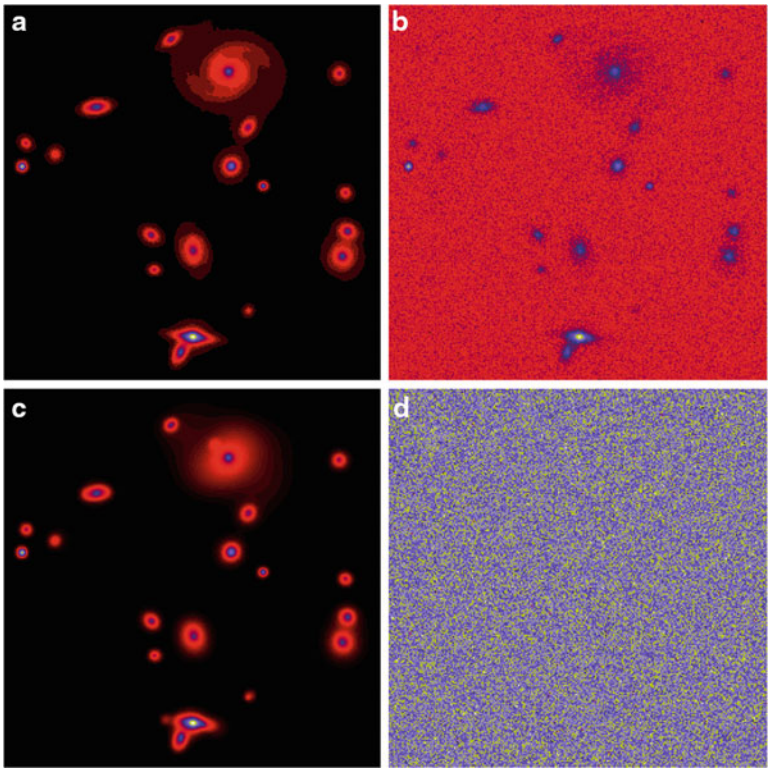


Fig. 8 (**a** and **b**) Original simulated image and the same image contaminated by a Gaussian noise. (**c** and **d**) Reconstructed image for the positive starlet coefficients of the noisy image using 50 iterations, and residual (i.e., noisy image – reconstructed image)

threshold t_j:

$$\tilde{w}_j[k,l] = \begin{cases} w_j[k,l] & \text{if } \mid w_j[k,l] \mid > t_j \\ 0 & \text{otherwise} \end{cases} \tag{27}$$

More generally, for a given sparse representation (i.e., wavelet) with its associated fast transform $\mathbf{W}$ and fast reconstruction $\mathbf{R}$, we can derive a hard threshold denoising solution X from the data Y, by first estimating the multiresolution support M using a given noise model, and then calculating

$$X = \mathbf{R}M\mathbf{W}Y. \tag{28}$$

We transform the data, multiply the coefficients by the support, and reconstruct the solution.

The solution can however be improved by considering the following optimization problem, $\min_X \parallel M(\mathbf{W}Y - \mathbf{W}X) \parallel_2^2$, where M is the multiresolution support of Y.

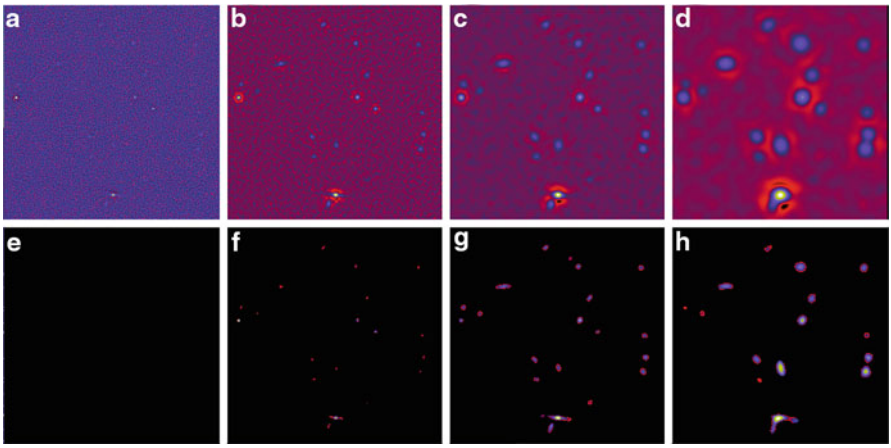


Fig. 9 *Top*, starlet transform, and *bottom*, positive starlet decomposition of a simulated astronomical image

A solution can be obtained using the Landweber iterative scheme [51, 58]:

$$X^{n+1} = X^n + \mathbf{R}M\left[\mathbf{W}Y - \mathbf{W}X^n\right] \tag{29}$$

If the solution is known to be positive, the positivity constraint can be introduced using the following equation:

$$X^{n+1} = P_+\left(X^n + \mathbf{R}M\left[\mathbf{W}Y - \mathbf{W}X^n\right]\right) \tag{30}$$

where P_+ is the projection on the cone of nonnegative images.

This algorithm allows us to constrain the residual to have a zero value within the multiresolution support [58]. For astronomical image filtering, iterating improves significantly the results, especially for the photometry (i.e., the integrated number of photons in a given object).

Removing the background in the solution is straightforward. The algorithm does not need to be modified. We only need to set to zero the coefficients related to the coarsest scale in the multiresolution support: $\forall k \quad M_J[k,l] = 0$.

The Multiscale Vision Model

Introduction

The wavelet transform of an image Y by the starlet transform produces at each scale j a set $\{w_j\}$. This has the same number of pixels as the image. The original image I can be expressed as the sum of all the wavelet scales and the smoothed array c_J by the expression

$$Y[k,l] = c_J[k,l] + \sum_{j=1}^{J} w_j[k,l]. \tag{31}$$

Hence, we have a *multiscale pixel representation*, i.e., each pixel of the input image is associated with a set of pixels of the multiscale transform. A further step is to consider a *multiscale object representation*, which would associate with an object contained in the data a volume in the multiscale transform. Such a representation obviously depends on the kind of image we need to analyze, and we present here a model which has been developed for astronomical data. It may however be used for other kinds of data, to the extent that such data are similar to astronomical data. We assume that an image Y can be decomposed into a set of components:

$$Y[k,l] = \sum_{i=1}^{N_o} X_i[k,l] + B[k,l] + N[k,l] \tag{32}$$

where N_o is the number of components, X_i are the components contained in the data (stars, galaxies, etc.), B is the background image, and N is the noise.

To perform such a decomposition, we have to detect, to extract, to measure, and to recognize the significant structures. This is done by first computing the multiresolution support of the image (i.e., the set of significant active coefficients) and by applying a segmentation scale by scale. The wavelet space of a 2D direct space is a 3D volume. An object, associated with a component, has to be defined in this space. A general idea for object definition lies in the connectivity property. An object occupies a physical region, and in this region we can join any pixel to other pixels based on significant adjacency. Connectivity in direct space has to be transported into wavelet transform space. In order to define the objects, we have to identify the wavelet transform space pixels we can attribute to the objects. We describe in this section the different steps of this method.

Multiscale Vision Model Definition

The multiscale vision model, MVM [8], described an object as a hierarchical set of structures. It uses the following definitions:

- *Significant wavelet coefficient*: a wavelet coefficient is significant when its absolute value is above a given detection limit. The detection limit depends on the noise model (Gaussian noise, Poisson noise, and so on). See section "Sparse Modeling of Astronomical Images" for more details.
- *Structure*: a structure $S_{j,k}$ is a set of significant connected wavelet coefficients at the same scale j:

$$S_{j,k} = \{w_j[k_1,l_1], w_j[k_2,l_2], \ldots, w_j[k_p,l_p]\} \tag{33}$$

where p is the number of significant coefficients included in the structure $S_{j,k}$ and $w_j[x_i, y_i]$ is a wavelet coefficient at scale j and at position (x_i, y_i).

- *Object*: an object is a set of structures:

$$O_l = \{S_{j_1,k_1}, \ldots, S_{j_n,k_n}\} \tag{34}$$

We define also the operator $\mathcal{L}$ which indicates to which object a given structure belongs: $\mathcal{L}(S_{j,k}) = l$ is $S_{j,k} \in O_l$, and $\mathcal{L}(S_{j,k}) = 0$ otherwise.

- *Object scale*: the scale of an object is given by the scale of the maximum of its wavelet coefficients.
- *Interscale relation*: the criterion allowing us to connect two structures into a single object is called the "interscale relation."
- *Sub-object*: a sub-object is a part of an object. It appears when an object has a local wavelet maximum. Hence, an object can be composed of several sub-objects. Each sub-object can also be analyzed.

From Wavelet Coefficients to Object Identification

Multiresolution Support Segmentation

Once the multiresolution support has been calculated, we have at each scale a Boolean image (i.e., pixel intensity equals 1 when a significant coefficient has been detected, and 0 otherwise). The segmentation consists of labeling the Boolean scales. Each group of connected pixels having a "1" value gets a label value between 1 and $L_{\max}$, $L_{\max}$ being the number of groups. This process is repeated at each scale of the multiresolution support. We define a "structure" $S_{j,i}$ as the group of connected significant pixels which has the label i at a given scale j.

Interscale Connectivity Graph

An object is described as a hierarchical set of structures. The rule which allows us to connect two structures into a single object is called "interscale relation." Figure 10 shows how several structures at different scales are linked together and form objects. We have now to define the interscale relation. Let us consider two structures at two successive scales, $S_{j,k}$ and $S_{j+1,l}$. Each structure is located in one of the individual images of the decomposition and corresponds to a region in this image where the signal is significant. Denoting (x_m, y_m) the pixel position of the maximum wavelet coefficient value of $S_{j,k}$, $S_{j,k}$ is said to be connected to $S_{j+1,l}$ if $S_{j+1,l}$ contains the pixel position (x_m, y_m) (i.e., the pixel position of the maximum wavelet coefficient of the structure $S_{j,k}$ must also be contained in the structure $S_{j+1,l}$). Several structures appearing in successive wavelet coefficient images can be connected in such a way, which we call an object in the interscale connectivity graph. Hence, we identify n_o objects in the wavelet space, each object O_i being defined by a set of structures, and we can assign to each structure a label i, with $i \in [1, n_o]$: $\mathcal{L}(S_{j,k}) = i$ if the structure $S_{j,k}$ belongs to the ith object.

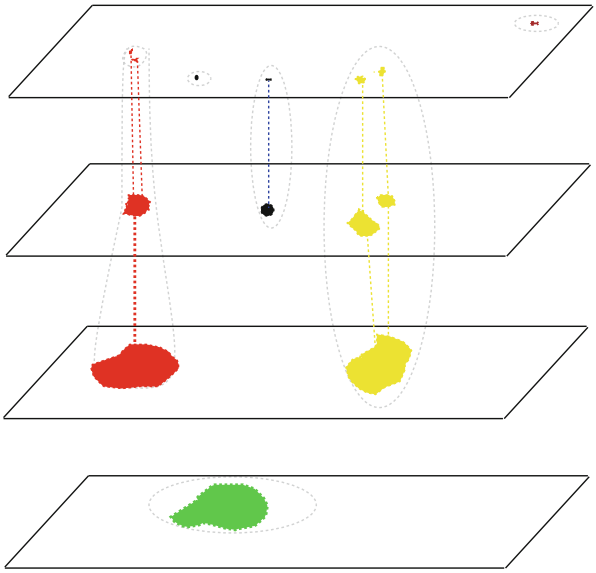


Fig. 10 Example of connectivity in wavelet space: contiguous significant wavelet coefficients form a structure, and following an interscale relation, a set of structures forms an object. Two structures S_j, S_{j+1} at two successive scales belong to the same object if the position pixel of the maximum wavelet coefficient value of S_j is included in S_{j+1}

Filtering

Statistically, some significant structures can be due to the noise. They contain very few pixels and are generally isolated, i.e., connected to no field at upper and lower scales. So, to avoid false detection, the isolated fields can be removed from the initial interscale connection graph. Structures at the border of the images may also have been detected because of the border problem and can be removed.

Merging/Deblending

As in the standard approach, true objects which are too close may generate a set of connected structures, initially associated with the same object, and a decision must be taken whether to consider such a case as one or two objects. Several cases may be distinguished:

- Two (or more) close objects, approximately of the same size, generate a set of structures. At a given scale j, two separate structures $S_{j,1}$ and $S_{j,2}$ are detected, while at the scale $j+1$, only one structure is detected $S_{j+1,1}$, which is connected to the $S_{j,1}$ and $S_{j,2}$.
- Two (or more) close objects of different sizes generate a set of structures, from scale j to scale k ($k > j$).

In the wavelet space, the merging/deblending decision will be based on the local maxima values of the different structures belonging to this object. A new object (i.e., deblending) is derived from the structure $S_{j,k}$ if there exists at least one other structure at the same scale belonging to the same object (i.e., there exists one structure $S_{j+1,a}$ and at least one structure $S_{j,b}$ such that $\mathcal{L}(S_{j+1,a}) = \mathcal{L}(S_{j,b}) = \mathcal{L}(S_{j,k})$) and if the following relationship is verified: $w_j^m > w_{j-1}^m$ and $w_j^m > w_{j+1}^m$, where:

- w_j^m is the maximum wavelet coefficient of the structure $S_{j,k}$: $w_j^m = \mathrm{Max}(S_{j,k})$:
  - $w_{j-1}^m = 0$ if $S_{j,k}$ is not connected to any structure at scale $j-1$.
  - w_{j-1}^m is the maximum wavelet coefficient of the structure $S_{j-1,l}$, where $S_{j-1,l}$ is such that $\mathcal{L}(S_{j-1,l}) = \mathcal{L}(S_{j,k})$ and the position of its highest wavelet coefficient is the closest to the position of the maximum of $S_{j,k}$.
- $w_{j+1}^m = \mathrm{Max}\{w_{j+1,x_1,y_1}, \ldots, w_{j+1,x_n,y_n}\}$, where all wavelet coefficients $w_{j+1,x,y}$ are at a position which belongs also to $S_{j,k}$ (i.e., $w_{j,x,y} \in S_{j,k}$).

When these conditions are verified, $S_{j,k}$ and all structures at smaller scales which are directly or indirectly connected to $S_{j,k}$ will define a new object.

Object Identification

We can now summarize this method allowing us to identify all the objects in a given image Y:

1. We compute the wavelet transform with the starlet algorithm, which leads to a set $\alpha = \mathbf{W}Y = \{w_1, \ldots, w_J, c_J\}$. Each scale w_j has the same size as the input image.
2. We determine the noise standard deviation in w_1.
3. We deduce the thresholds at each scale from the noise modeling.
4. We threshold scale by scale and we do an image labeling.
5. We determine the interscale relations.
6. We identify all the wavelet coefficient maxima of the wavelet transform space.
7. We extract all the connected trees resulting from each wavelet transform space maximum.

Source Reconstruction

Partial Reconstruction as an Inverse Problem

A set of structures $\mathcal{S}_i$ ($\mathcal{S}_i = \{S_{j,k}, \ldots, S_{j',k'}\}$) defines an object O_i which can be reconstructed separately from other objects, in order to provide the components X_i. The co-addition of all reconstructed objects is a filtered version of the input data. We will denote α_i the set of wavelet coefficients belonging to the object O_i. Therefore, α_i is a subset of the wavelet transform of X_i, $\tilde{\alpha}_i = \mathbf{W}X_i$. Indeed, the last scale of

$\tilde{\alpha}_i$ is unknown, as well as many wavelet coefficients which have not been detected. Then the reconstruction problem consists of searching for an image X_i such that its wavelet transform reproduces the coefficients α_i (i.e., they are the same as those of $\mathcal{S}_i$, the detected structures). If $\mathbf{W}$ describes the wavelet transform operator and P_w the projection operator in the subspace of the detected coefficients (i.e., having set to zero all coefficients at scales and positions where nothing was detected), the solution is found by minimization of

$$\min_{X_i} \| \alpha_i - P_w (\mathbf{W} X_i) \|^2 \tag{35}$$

The size of the restored image X_i is arbitrary, and it can be easily set greater than the number of known coefficients. It is certain that there exists at least one image X_i which gives exactly α_i, i.e., the original one. But generally we have an infinity of solutions, and we have to choose among them the one which is considered as correct. An image is always a positive function, which leads us to constrain the solution, but this is not sufficient to get a unique solution. More details on the reconstruction algorithm can be found in [8, 57].

Examples

Band Extraction

We simulated a spectrum which contains an emission band at 3.50 μm and a nonstationary noise superimposed on a smooth continuum. The band is a Gaussian of width FWHM = 0.01 μm (FWHM = full width at half-maximum) and normalized such that its maximum value equals ten times the local noise standard deviation.

Figure 11 (top) contains the simulated spectrum. The wavelet analysis results in the detection of an emission band at 3.50 μm above 3σ. Figure 11 (middle) shows the reconstruction of the detected band in the simulated spectrum. The real feature is overplotted as a dashed line. Figure 11 (bottom) contains the original simulation with the reconstructed band subtracted. It can be seen that there are no strong residuals near the location of the band, which indicates that the band is well reconstructed. The center position of the band, its FWHM, and its maximum can then be estimated via a Gaussian fit. More details about the use of MVM for spectral analysis can be found in [60].

Star Extraction in NGC 2997

We applied MVM to the galaxy NGC 2997 (Fig. 12, top left). Two images were created by co-adding objects detected from scales 1 and 2 and from scales 3 to 6. They are displayed, respectively, in Fig. 12, top right and bottom left. Figure 12, bottom right, shows the difference between the input data and the image which contained the objects from scales 1 and 2. As we can see, all small objects have been removed, and the galaxy can be better analyzed.

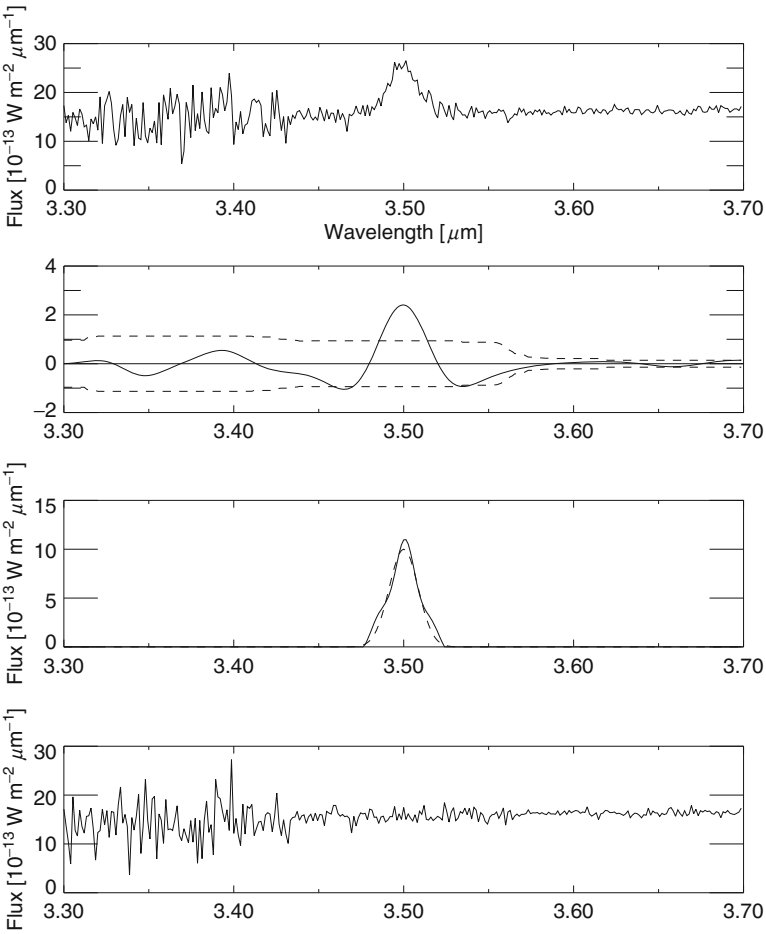


Fig. 11 *Top*: simulated spectrum. *Middle*: reconstructed simulated band (*full line*) and original band (*dashed line*). *Bottom*: simulated spectrum minus the reconstructed band

Galaxy Nucleus Extraction

Figure 13 shows the extracted nucleus of NGC 2997 using the MVM method and the difference between the galaxy image and the nucleus image.

5 Deconvolution

Up to now, the PSF H has not been considered in the source detection. This means that all morphological parameters (size, ellipticity, etc.) derived from the detected objects need to be corrected from the PSF. Very close objects may also be seen as a single object because H acts as a blurring operator on the data. A solution

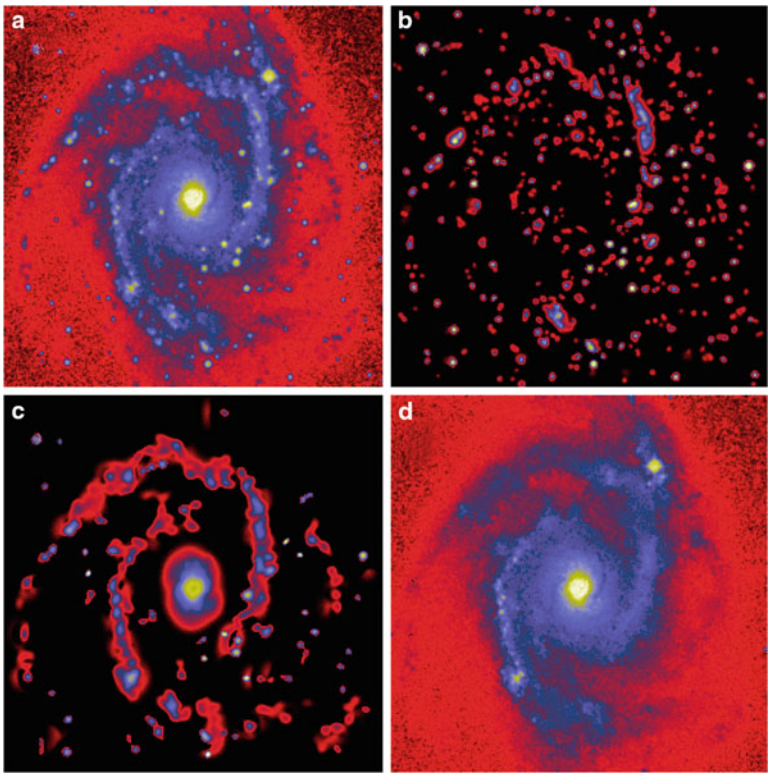


Fig. 12 (**a**) Galaxy NGC 2997, (**b**) objects detected from scales 1 and 2, (**c**) objects detected from scales 3 to 6, and (**d**) difference between (**a**) and (**b**)

may consist in deconvolving first the data and carrying out the source detection afterwards.

The problem of image deconvolution is ill-posed [3], and as a consequence, the matrix H modeling the imaging system is ill-conditioned. If Y is the observed image and X the unknown object, the equation $HX = Y$ has not a unique and stable solution. Therefore, one must look for approximate solutions of this equation that are also physically meaningful. One approach is Tikhonov regularization theory [23]; however, a more general approach is provided by the so-called Bayes paradigm [25], even if it is applicable only to discrete problems. In this framework one can both take into account statistical properties of the data (Tikhonov regularization is obtained by assuming additive Gaussian noise) and also introduce a priori information on the unknown object.

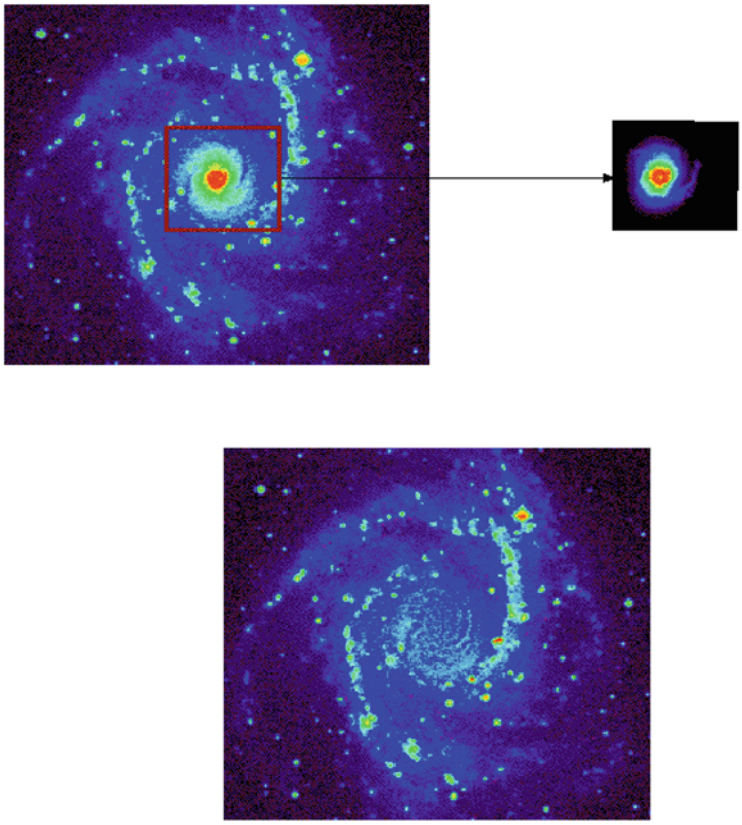


Fig. 13 *Upper left*, galaxy NGC 2997; *upper right*, extracted nucleus; *bottom*, difference between the two previous images

Statistical Approach to Deconvolution

We assume that the detected image Y is the realization of a multivalued random variable I corresponding to the (unknown) value X of another multivalued random variable, the object O. Moreover, we assume that the *conditional probability distribution* $p_I(Y|X)$ is known. Since the unknown object appears as a set of unknown parameters, the problem of image deconvolution can be considered as a classical problem of parameter estimation. The standard approach is the *maximum likelihood* (ML) method. In our specific application, for a given detected image Y, this consists of introducing the *likelihood function* defined by

$$L_Y(X) = p_I(Y;X) \ . \tag{36}$$

Then the ML estimate of the unknown object is any maximizer X^* of the likelihood function

$$X^* = \arg\max_{X \in \mathbb{R}^n} L_Y(X) \ , \tag{37}$$

if it exists.

In our applications the likelihood function is the product of a very large number of terms (the data components are assumed to be statistically independent), so that it is convenient to take the logarithm of this function; moreover, if we consider its negative logarithm, the maximization problem is transformed into a minimization one. Let us consider the function

$$J_0(X;Y) = -A \ln L_Y(X) + B \ , \tag{38}$$

where A, B are suitable constants. They are introduced in order to obtain a function which has a simple expression and is also nonnegative since, in our applications, the negative logarithm of the likelihood is bounded from below. Then, it is easy to verify that the problem of Eq. 37 is equivalent to the following one:

$$X^* = \arg\min_{X \in \mathbb{R}^n} J_0(X;Y) \ . \tag{39}$$

We consider now the model of Eq. 2 with three different examples of noise.

Example 1. In the case of additive white Gaussian noise, by a suitable choice of the constants A, B, we obtain (we assume here that the background B is not subtracted even if it must be estimated)

$$J_0(X;Y) = ||HX + B - Y||^2 \ , \tag{40}$$

and therefore, the ML approach coincides with the well-known least-squares (LS) approach. It is also well known that the function of Eq. 40 is convex and strictly convex if and only if the equation $HX = 0$ has only the solution $X = 0$. Moreover, it has always absolute minimizers, i.e., the LS problem has always a solution; but the problem is ill-conditioned because it is equivalent to the solution of the Euler equation:

$$H^T H \ X = H^T (Y - B) \ . \tag{41}$$

We remark that the ill-posedness of the LS problem is the starting point of *Tikhonov regularization theory* (see, for instance, [23,63]), and therefore, this theory is based on the tacit assumption that the noise affecting the data is additive and Gaussian.

We remark that, in the case of object reconstruction, since objects are non-negative, we should consider the minimization of the function of Eq. 40 on the

nonnegative orthant. With such a constraint the problem is not treatable in the standard framework of regularization theory.

Example 2. In the case of Poisson noise, if we introduce the so-called generalized Kullback–Leibler (KL) divergence of a vector Z from a vector Y, defined by

$$D_{\mathrm{KL}}(Y,Z)=\sum_{i=1}^{m}\left\{ y_i \ln\frac{Y_i}{Z_i}+Z_i-Y_i\right\} \quad , \tag{42}$$

then, with a suitable choice of the constants A, B, the function $J_0(X;Y)$ is given by

$$\begin{aligned} J_0(X;Y) &= D_{\mathrm{KL}}(Y;HX+B) \\ &= \sum_{i=1}^{m}\left\{ Y_i \ln\frac{y_i}{(HX+B)_i}+(HX+B)_i-y_i\right\} \quad . \end{aligned} \tag{43}$$

It is quite natural to take the nonnegative orthant as the domain of this function. Moreover, it is well known that it is convex (strictly convex if the equation $HX=0$ has only the solution $X=0$), nonnegative, and coercive. Therefore, it has absolute minimizers. However, these minimizers are strongly affected by noise, and the specific effect of the noise in this problem is known as *checkerboard effect* [41], since many components of the minimizers are zero.

Example 3. In the case of Gauss+Poisson noise, the function $J_0(X;Y)$ is given by a much more complex form. This function is also convex (strictly convex if the equation $Hx=0$ has the unique solution $x=0$), nonnegative, and coercive [2]. Therefore, it also has absolute minimizer on the nonnegative orthant.

The previous examples demonstrate that, in the case of image reconstruction, ML problems are ill-posed or ill-conditioned. That means that one is not interested in computing the minimum points X^* of the functions corresponding to the different noise models because they do not provide sensible estimates $\bar{X}$ of the unknown object.

The previous remark is not surprising in the framework of inverse problem theory. Indeed it is generally accepted that, if the formulation of the problem does not use some additional information on the object, then the resulting problem is ill-posed. This is what happens in the maximum likelihood approach because we only use information about the noise with, possibly, the addition of the constraint of non-negativity.

The additional information may consist, for instance, of prescribed bounds on the solution and/or its derivatives up to a certain order (in general not greater than two). These prescribed bounds can be introduced in the problem as additional constraints in the variational formulation provided by ML. However, in a quite natural

probabilistic approach, called the *Bayesian approach*, the additional information is given in the form of statistical properties of the object [25].

In other words, one assumes that the unknown object X is a realization of a vector-valued random variable O and that the probability distribution of O, the so-called prior denoted by $p_O(X)$, is also known or can be deduced from known properties of the object. The most frequently used priors are Markov random fields or, equivalently, Gibbs random fields, i.e., they have the following form:

$$p_O(X) = \frac{1}{Z} \mathrm{e}^{-\mu \Omega(X)} \ , \tag{44}$$

where Z is a normalization constant, μ is a positive parameter (a hyperparameter in statistical language, a regularization parameter in the language of regularization theory), while $\Omega(X)$ is a function, possibly convex.

The previous assumptions imply that the joint probability density of the random variables O, I is given by

$$p_{OI}(X, Y) = p_I(Y|X) p_O(X) \ . \tag{45}$$

If we introduce the marginal probability density of the image I

$$p_I(Y) = \int p_{OI}(X, Y) \, dX \ , \tag{46}$$

from *Bayes' formula* we obtain the conditional probability density of O for a given value Y of I:

$$p_O(X|Y) = \frac{p_{OI}(X, Y)}{p_I(Y)} = \frac{p_I(Y|X) p_O(X)}{p_I(Y)} \ . \tag{47}$$

If in this equation we insert the detected value Y of the image, we obtain the a posteriori probability density of X:

$$P_Y(X) = p_O(X|Y) = L_Y(X) \frac{p_O(X)}{p_I(Y)} \ . \tag{48}$$

Then, a *maximum a posteriori* (MAP) estimate of the unknown object is defined as any object X^* that maximizes the a posteriori probability density:

$$X^* = \arg\max_{X \in \mathbb{R}^n} P_Y(X) \ . \tag{49}$$

As in the case of the likelihood, it is convenient to consider the negative logarithm of $P_Y(X)$. If we assume a Gibbs prior as that given in Eq. 44 and we take into account the definition of Eq. 38, we can introduce the following function:

$$\begin{aligned} J(X;Y) &= -A\ln P_Y(X) + B - A\ln Z \\ &- A\ln p_I(Y) = J_0(X;Y) + \mu J_R(X)\ , \end{aligned} \tag{50}$$

where $J_R(X) = A\Omega(X)$. Therefore, the MAP estimates are also given by

$$X^* = \arg\min_{X\in\mathbb{R}^n} J(X;Y) \tag{51}$$

and again one must look for the minimizers satisfying the non-negativity constraint.

The Richardson–Lucy Algorithm

One of the most frequently used methods for image deconvolution in astronomy is an iterative algorithm known as the *Richardson–Lucy* (RL) algorithm [34, 48]. In emission tomography it is also denoted as *expectation maximization* (EM) because, as shown in [49], it can be obtained by applying to the ML problem with Poisson noise a general EM method introduced in [19] for obtaining ML estimates.

In [49] it is shown that, if the iteration converges, then the limit is just an ML estimate in the case of Poisson data. Subsequently the convergence of the algorithm was proved by several authors in the case $B = 0$. An account can be found in [41].

The iteration is as follows: it is initialized with a positive image $X^{(0)}$ (a constant array, in general); then, given $X^{(n)}$, $X^{(n+1)}$ is computed by

$$X^{(n+1)} = X^{(n)} H^T \frac{Y}{HX^{(n)} + B}\ . \tag{52}$$

This algorithm has some nice features. First, the result of each iteration is automatically a positive array; second, in the case $B = 0$, the result of each iteration has the same flux of the detected image Y, and this property is interesting from the photometric point of view.

The limit of the RL iteration is, in general, very noisy and sparse in pixel space (see the remark at the end of Example 2 in the previous section) and can provide satisfactory results in the case of star systems (see [4], section 3.1); in the case of complex systems, a reasonable solution can be obtained by a suitable stopping of the algorithm before convergence. This can be seen as a kind of regularization, and this property is called *semi-convergence* [3], i.e., the iteration first approaches the correct solution and then goes away. An example of RL reconstruction is shown in Fig. 14 (lower left panel).

The main drawback of RL is that, in general, it is very slow and may require hundreds or thousands of iterations. The proposed acceleration approaches are based on the remark that RL is a scaled gradient method since it can be written in the

following form:

$$X^{(n+1)} = X^{(n)} - X^{(n)} \nabla_X J_0(X^{(n)}; Y) \ , \tag{53}$$

where ∇_X denotes the gradient with respect to X and $J_0(X; Y)$ is the data-fidelity function defined in Eq. 43. Therefore, a reduction of the number of iterations can be obtained by means of a suitable line search along the descent direction. This is the approach proposed by several authors. However, this structure of RL has inspired a recently proposed optimization method, known as *scaled gradient projection* (SGP) [10], which can be viewed as a general method for the constrained minimization of differentiable functions.

If $J(X)$ is the function to be minimized on the nonnegative orthant, then the method is based on the descent direction:

$$D^{(n)} = P_+\left(X^{(n)} - \sigma_n S^{(n)} \nabla_X J(X^{(n)})\right) - X^{(n)} \tag{54}$$

where P_+ is the projection on the nonnegative orthant, $S^{(n)}$ is the (diagonal) scaling matrix, and σ_n is a suitably chosen step length [10]. Then iteration $X^{(n+1)}$ is obtained by a line search along the descent direction based on the Armijo rule. We can add that the method can be easily extended to the case where the convex set of the admissible solutions is defined by box and equality constraints.

In the case of the minimization of the KL divergence, the diagonal scaling matrix is that suggested by RL, i.e., $S^{(n)} = X^{(n)}$ (with the addition of suitable upper and lower bounds), and the method shows the semi-convergence property as RL, but requires a much smaller number of iterations for obtaining a sensible reconstruction. In an application to the deconvolution of astronomical images [46], it has been shown that, even if the computational cost per iteration is about 30 % greater than that of RL, thanks to the reduction of the number of iterations it is possible to obtain a speedup, with respect to RL, ranging from 4 to more than 30, depending on the astronomical source and the noise level. Moreover, implementation on *graphics processing units* (GPU) allows to deconvolve a 2,048 × 2,048 image in a few seconds.

Several iterative methods, modeled on RL, have been introduced for computing MAP estimates corresponding to different kinds of priors. A recent account can be found in [4]. A nice feature of SGP, whose convergence is proved in [10], is that it can be easily applied to this problem, i.e., to the minimization of the function of Eq. 50 (again, with the addition of box and equality constraints). The scaling is taken from the *split-gradient method* (SGM), proposed in [31], since this scaling is always nonnegative, while the scaling proposed in [26] may take negative values. In general the choice of the scaling is as follows. If the gradient of $J_R(X)$ is split in the following way:

$$-\nabla_X J_R(X) = U_R(X) - V_R(X), \tag{55}$$

where U_R, V_R are nonnegative arrays, then the scaling is given by the array

$$S = \frac{X}{1 + \mu V_R(X)} \ . \tag{56}$$

Of course SGP can be applied if $J_R(X)$ is differentiable, and therefore, it can cover both smoothing regularization as given by Tikhonov and also edge-preserving regularization as given by smoothed TV (total variation) [12, 65]. Finally, a difficult problem in the case of regularized problems is the choice of the value of the regularization parameter. An attempt in the direction of solving this problem is provided by a recently proposed discrepancy principle for Poisson data deconvolution [5].

Blind Deconvolution

Blind deconvolution is the problem of image deblurring when the blur is unknown. In the case of a space-invariant model, the naive problem formulation is to solve the problem $Y = H * X$ where only Y is known, where $*$ denotes convolution. It is obvious that this problem is extremely undetermined and that there is an infinite set of pairs solving the equation. Among them also is the trivial solution $X = Y$, $H = \delta$, where δ denotes the usual delta function. Therefore, the problem must be formulated by introducing as far as possible all available constraints on both the object X and the PSF H.

In the case of Poisson noise, several iterative methods have been proposed, which consist of alternating updates of the object and PSF by means of RL iterations or accelerated RL iterations. For instance, in [27] one RL iteration is used both on the object and the PSF. This algorithm was investigated, in the context of *nonnegative matrix factorization* (NMF), by Lee and Seung [32], but their convergence proof is incomplete, since only the monotonic decrease of the objective function is shown while, for a general descent method to be convergent, strongest Armijo-like decreasing conditions have to be verified. In general, the proposed approaches to blind deconvolution with Poisson data could be classified as methods of inexact alternating minimization applied to the KL divergence, as a function of both object and PSF.

In a recent paper [9], in the context of NMF, convergence of inexact alternating minimization is proved if the iterative algorithm used for the inner iterations satisfies suitable conditions, which are satisfied by SGP. Therefore, this approach looks very suitable for the problem of blind deconvolution with Poisson data. The approach is applied in [47] using constraints both on the object and the PSF. The method applies to the imaging by ground-based telescopes since, as suggested in [20], one of the constraints on the PSF is provided by the Strehl ratio (SR; see Sect. 2), a parameter measuring the quality of the AO correction. Indeed we recall that the advantage of

SGP is not only fast convergence, if a suitable scaling of the gradient is used, but also the possibility of introducing suitable box and equality constraints on the solution. In particular the SR constraint excludes the trivial solution mentioned above. The method works well in the case of star systems.

Deconvolution with a Sparsity Prior

Another approach is to use the sparsity to model the data. A sparse model can be interpreted from a Bayesian standpoint, by assuming the coefficients α of the solution in the dictionary Φ follow a leptokurtic PDF with heavy tails such as the generalized Gaussian distribution form:

$$\mathrm{pdf}_{\alpha}(\alpha_1, \ldots, \alpha_K) \propto \prod_{k=1}^{K} \exp\left(-\lambda \left\|\alpha_i\right\|_p^p\right) \quad 0 \leq p < 2\,. \tag{57}$$

Between all possible solutions, we want the one which has the sparsest representation in the dictionary Φ. Putting together the log-likelihood function in the case of Gaussian noise σ and the priors on α, the MAP estimator leads to the following optimization problem:

$$\min_{\alpha_1,\ldots,\alpha_K} \frac{1}{2\sigma} \left\|Y - \Phi\alpha\right\|^2 + \lambda \sum_{k=1}^{K} \left\|\alpha_k\right\|_p^p \,, 0 \leq p < 2\,. \tag{58}$$

The sparsity can be measured through the $\|\alpha\|_0$ norm (i.e., $p = 0$). This counts in fact the number of nonzero elements in the sequence. It was also proposed to convexify the constraint by substituting the convex $\|\alpha\|_1$ norm for the $\|\alpha\|_0$ norm [13]. Depending on the H operator, there are several ways to obtain the solution of this equation.

A first iterative thresholding deconvolution method was proposed in [51] which consists of the following iterative scheme:

$$X^{(n+1)} = P_+\left(X^{(n)} + H^T\left(\mathbf{WDen}_{M^{(n)}}\left(Y - HX^{(n)}\right)\right)\right) \tag{59}$$

where P_+ is the projection on the cone of nonnegative images and **WDen** is an operator which performs a wavelet thresholding, i.e., applies the wavelet transform of the residual $R^{(n)}$ (i.e., $R^{(n)} = Y - HX^{(n)}$), thresholds some wavelet coefficients, and applies the inverse wavelet transform. Only coefficients that belong to the multiresolution support $M^{(n)}$ [51] are kept, while the others are set to zero. At each iteration, the multiresolution support $M^{(n)}$ is updated by selecting new coefficients in the wavelet transform of the residual which have an absolute value larger than a given threshold. The threshold is automatically derived assuming a given noise distribution such as Gaussian or Poisson noise.

More recently, it was shown [14, 16, 24] that a solution of Eq. 58 for $p = 1$ can be obtained through a thresholded Landweber iteration:

$$X^{(n+1)} = P_+ \left(\mathbf{WDen}_\lambda \left(X^{(n)} + H^T \left(Y - HX^{(n)} \right) \right) \right) , \tag{60}$$

with $\|H\| = 1$. In the framework of monotone operator splitting theory, it was shown that for frame dictionaries, a slight modification of this algorithm converges to the solution [14]. Extension to constrained nonlinear deconvolution is proposed in [22].

Constraints in the Object or Image Domains

Let us define the *object domain* $\mathcal{O}$ as the space in which the solution belongs and the *image domain* $\mathcal{I}$ as the space in which the observed data belongs (i.e., if $X \in \mathcal{O}$ then $HX \in \mathcal{I}$). The constraint in (59) was applied in the image domain, while in (60) we have considered constraints on the solution. Hence, two different wavelet-based strategies can be chosen in order to regularize the deconvolution problem. The constraint in the image domain through the multiresolution support leads to a very robust way to control the noise. Indeed, whatever the nature of the noise, we can always derive robust detection levels in the wavelet space and determine scales and positions of the important coefficients. A drawback of the image constraints is that there is no guarantee that the solution is free of artifacts such as ringing around point sources. A second drawback is that image constraints can be used only if the point spread function is relatively compact, i.e., does not smear the information over the whole image.

The property of introducing robust noise modeling is lost when applying the constraint in the object domain. For example, in the case of Poisson noise, there is no way (except using time-consuming Monte Carlo techniques) to estimate the level of the noise in the solution and to adjust properly the thresholds. The second problem with this approach is that, in fact, we try to solve two problems simultaneously (noise amplification and artifact control in the solution) with one parameter (i.e., λ). The choice of this parameter is crucial, while such a parameter is implicit when using the multiresolution support.

Ideally, constraints should be added in both the object and image domains in order to better control the noise by using the multiresolution support and avoid such a ringing artifact.

Example

A simulated Hubble Space Telescope Wide Field Camera image of a distant cluster of galaxies is shown in Fig. 14, upper left. The simulated data are shown in Fig. 14, upper right. The Richardson–Lucy and the wavelet solutions are shown, respectively, in Fig. 14, lower left and right. The Richardson–Lucy method amplifies the noise, which implies that the faintest objects disappear in the deconvolved image, while the wavelet starlet solution is stable for any kind of PSF, and any kind of noise modeling can be considered.

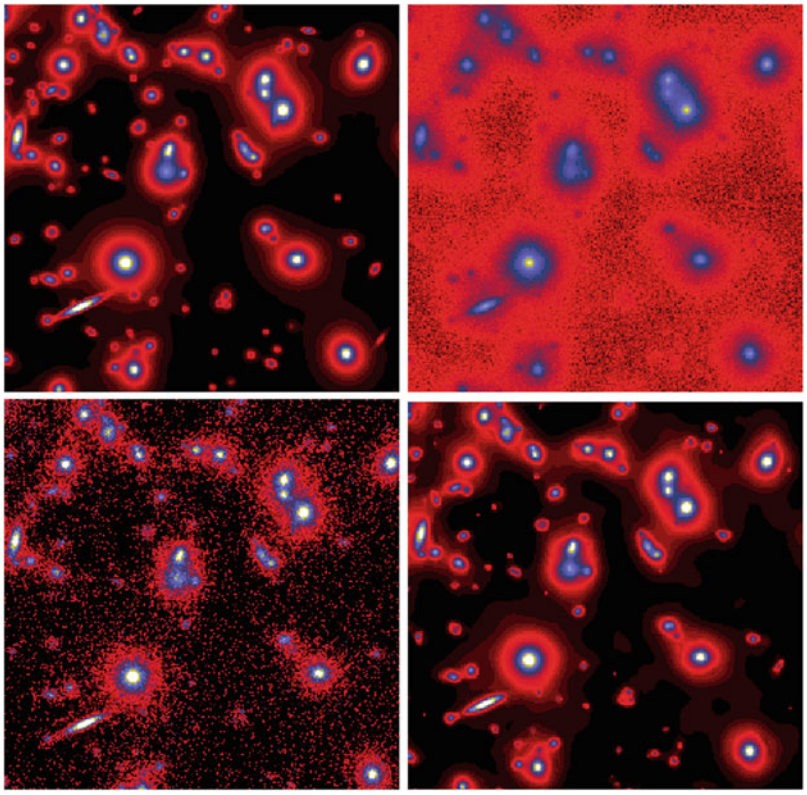


Fig. 14 Simulated Hubble Space Telescope Wide Field Camera image of a distant cluster of galaxies. *Upper left*: original, unaberrated, and noise-free. *Upper right*: input, aberrated, noise added. *Lower left*: restoration, Richardson–Lucy. *Lower right*, restoration starlet deconvolution

Detection and Deconvolution

The PSF is not needed with MVM. This is an advantage when the PSF is unknown or difficult to estimate, which happens relatively often when it is space variant. However, when the PSF is well determined, it becomes a drawback because known information is not used for the object reconstruction. This can lead to systematic errors in the photometry, which depends on the PSF and on the source signal-to-noise ratio. In order to preempt such a bias, a kind of calibration must be performed using simulations [50]. This section shows how the PSF can be used in the MVM, leading to a deconvolution.

Object Reconstruction Using the PSF

A reconstructed and deconvolved object X_i can be obtained by searching for a signal X_i such that the wavelet coefficients of HX_i are the same as those of the

detected structures α_i. If $\mathbf{W}$ describes the wavelet transform operator and P_w the projection operator in the subspace of the detected coefficients, the solution is found by minimization of

$$\min_{X_i} \| \alpha_i - P_w (\mathbf{W} H X_i) \|^2 \tag{61}$$

where α_i represents the detected wavelet coefficients of the object O_i and H is the PSF. In this approach, each object is deconvolved separately. The flux related to the extent of the PSF will be taken into account. For point sources, the solution will be close to that obtained by PSF fitting. This problem is also different from global deconvolution in the sense that it is well constrained. Except for the positivity of the solution which is always true and must be used, no other constraint is needed. This is due to the fact that the reconstruction is performed from a small set of wavelet coefficients (those above a detection limit). The number of objects is the same as those obtained by the MVM, but the photometry and the morphology are different. The astrometry may also be affected.

The Algorithm

Any minimizing method can be used to obtain the solution X_i. Since there is no problem of convergence, noise amplification, or ringing effect, the Van Cittert method was proposed on the grounds of its simplicity [57]. It leads to the following iterative scheme:

$$X_i^{(n+1)} = X_i^{(n)} + \mathbf{R}\left(\alpha_i - P_w\left(\mathbf{W} H X_i^{(n)}\right)\right) \tag{62}$$

where $\mathbf{R}$ is the inverse wavelet transform, and the algorithm is:

1. Set n to 0.
2. Find the initial estimation $X_i^{(n)}$ by applying an inverse wavelet transform to the set α_i corresponding to the detected wavelet coefficients in the data.
3. Convolve $X_i^{(n)}$ with the PSF H: $Y_i^{(n)} = H X_i^{(n)}$.
4. Determine the wavelet transform $\alpha^{(Y_i^{(n)})}$ of $Y_i^{(n)}$.
5. Threshold all wavelet coefficients in $\alpha^{(Y_i^{(n)})}$ at position and scales where nothing has been detected (i.e., P_w operator). We get $\alpha_t^{(Y_i^{(n)})}$.
6. Determine the residual $\alpha_r = \alpha_i - \alpha_t^{(Y_i^{(n)})}$.
7. Reconstruct the residual image $R^{(n)}$ by applying an inverse wavelet transform.
8. Add the residual to the solution: $X_i^{(n+1)} = X_i^{(n)} + R^{(n)}$.
9. Threshold negative values in $X_i^{(n+1)}$.
10. If $\sigma(R^{(n)})/\sigma(X_i^{(0)}) < \epsilon$, then $n = n + 1$ and go to step 3.
11. $X_i^{(n+1)}$ contains the deconvolved reconstructed object.

In practice, convergence is very fast (less than 20 iterations). The reconstructed image (not deconvolved) can also be obtained just by reconvolving the solution with the PSF.

Space-Variant PSF

Deconvolution methods generally do not take into account the case of a space-variant PSF. The standard approach when the PSF varies is to decompose the image into blocks and to consider the PSF constant inside a given block. Blocks which are too small lead to a problem of computation time (the FFT cannot be used), while blocks which are too large introduce errors due to the use of an incorrect PSF. Blocking artifacts may also appear. Combining source detection and deconvolution opens up an elegant way for deconvolution with a space-variant PSF. Indeed, a straightforward method is derived by just replacing the constant PSF at step 3 of the algorithm with the PSF at the center of the object. This means that it is not the image which is deconvolved, but its constituent objects.

Undersampled Point Spread Function

If the PSF is undersampled, it can be used in the same way, but results may not be optimal due to the fact that the sampled PSF varies depending on the position of the source. If an oversampled PSF is available, resulting from theoretical calculation or from a set of observations, it should be used to improve the solution. In this case, each reconstructed object will be oversampled. Equation 61 must be replaced by

$$\min_{X_i} \parallel \alpha_i - P_w \left(\mathbf{W}\mathcal{D}_l H X_i\right) \parallel^2 \tag{63}$$

where $\mathcal{D}_l$ is the averaging-decimation operator, consisting of averaging the data in the window of size $l \times l$ and keeping only one average pixel for each $l \times l$ block.

Example: Application to Abell 1689 ISOCAM Data

Figure 15 (left) shows the detections (isophotes) obtained using the MVM method without deconvolution on ISOCAM data. The data were collected using the 6 arcsec lens at 6.75 μm. This was a raster observation with 10 s integration time, 16 raster positions, and 25 frames per raster position. The noise is nonstationary, and the detection of the significant wavelet coefficients was carried out using the root mean square error map $R_\sigma(x, y)$ by the method described in [50]. The isophotes are overplotted on an optical image (NTT, band V) in order to identify the infrared source. Figure 15 (right) shows the same treatment but using the MVM method with deconvolution. The objects are the same, but the photometry is improved, and it is clearly easier to identify the optical counterpart of the infrared sources.

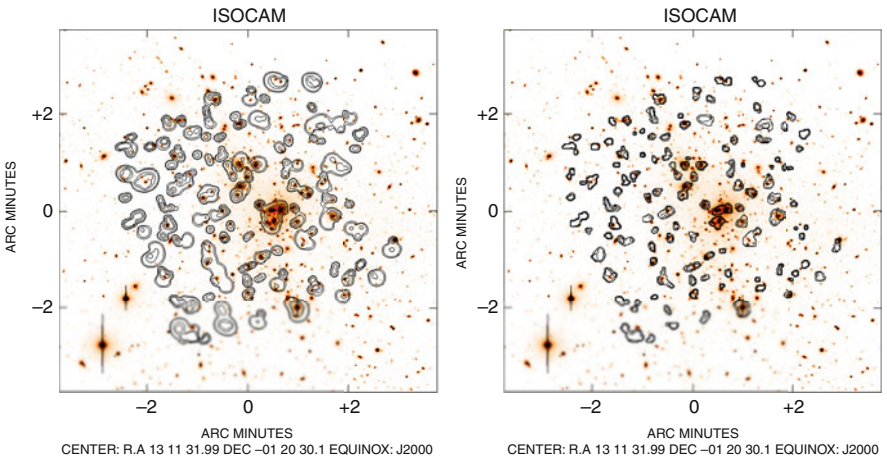


Fig. 15 Abell 1689: *left*, ISOCAM source detection (isophotes) overplotted on an optical image (NTT, band V). The ISOCAM image is a raster observation at 7 μm. *Right*, ISOCAM source detection using the PSF (isophotes) overplotted on the optical image. Compared to the *left panel*, it is clearly easier to identify the detected infrared sources in the optical image

6 Conclusion

In this chapter we have used the sparsity principle that now occupies a very central role in signal processing. We have discussed the vision models within which the sparsity principle is applied. Finally, we have reviewed the use of the starlet wavelet transform as a prime technique in order to apply the sparsity principle in the context of vision models in various application domains. Among the latter are object detection coupled with denoising, deconvolution and ltering generally. Issues of algorithmic optimization and of statistical modeling entered into our discussion on various occasions. Many examples and case studies were used to demonstrate the powerfulness of the approaches described for astronomical data analysis and processing.

Cross-References

- ▸ Energy Minimization Methods
- ▸ Iterative Solution Methods
- ▸ Large-Scale Inverse Problems in Imaging
- ▸ Linear Inverse Problems
- ▸ Numerical Methods and Applications in Total Variation Image Restoration
- ▸ Regularization Methods for Ill-Posed Problems
- ▸ Splines and Multiresolution Analysis
- ▸ Statistical Methods in Imaging
- ▸ Total Variation in Imaging

Recommended Readings

- Bertero, M., Boccacci, P.: Introduction to Inverse Problems in Imaging. Institute of Physics (1998)
- Starck, J.-L., Murtagh, F.: Astronomical Data Analysis, 2nd edn. Springer (2006).
- Mallat, S.: A Wavelet Tour of Signal Processing, 3rd edn. Academic (2008).
- Starck, J.-L., Murtagh, F., Fadili, J.: Sparse Image & Signal Processing. Cambridge University Press (2010)

References

1. Anscombe, F.J.: The transformation of Poisson, binomial and negative-binomial data. Biometrika **15**, 246–254 (1948)
2. Benvenuto, F., La Camera, A., Theys, C., Ferrari, A., Lantéri, H., Bertero, M.: The study of an iterative method for the reconstruction of images corrupted by Poisson and Gaussian noise. Inverse Probl. **24**(035016), 20pp (2008)
3. Bertero, M., Boccacci, P.: Introduction to Inverse Problems in Imaging. Institute of Physics, Bristol/Philadelphia (1998)
4. Bertero, M., Boccacci, P., Desiderá, G., Vicidomini, G.: Image deblurring with Poisson data: from cells to galaxies. Inverse Probl. **25**(123006), 26pp (2009)
5. Bertero, M., Boccacci, P., Talenti, G., Zanella, R., Zanni, L.: A discrepancy principle for Poisson data. Inverse Probl. **26**, 10500 (2010)
6. Bertin, E., Arnouts, S.: SExtractor: software for source extraction. Astron. Astrophys. Suppl. Ser. **117**, 393–404 (1996)
7. Bijaoui, A.: Sky background estimation and application. Astron. Astrophys. **84**, 81–84 (1980)
8. Bijaoui, A., Rué, F.: A multiscale vision model adapted to astronomical images. Signal Process. **46**, 229–243 (1995)
9. Bonettini, S.: Inexact block coordinate descent methods with application to non-negative matrix factorization. IMA J. Numer. Anal. **31**(4), 1431–1452 (2011)
10. Bonettini, S., Zanella, R., Zanni, L.: A scaled gradient projection method for constrained image deblurring. Inverse Probl. **25**(1), 015002 (2009)
11. Buonanno, R., Buscema, G., Corsi, C.E., Ferraro, I., Iannicola, G.: Automated photographic photometry of stars in globular clusters. Astron. Astrophys. **126**, 278–282 (1983)
12. Charbonnier, P., Blanc-Féraud, L., Aubert, G., Barlaud, M.: Deterministic edge-preserving regularization in computed imaging. IEEE Trans. Image Process. **6**, 298–311 (1997)
13. Chen, S.S., Donoho, D.L., Saunders, M.A.: Atomic decomposition by basis pursuit. SIAM J. Sci. Comput. **20**(1), 33–61 (1999)
14. Combettes, P.L., Wajs, V.R.: Signal recovery by proximal forward-backward splitting. Multiscale Model. Simul. **4**(4), 1168–1200 (2005)
15. Da Costa, G.S.: Basic photometry techniques. In: Howel, S.B. (ed.) Astronomical CCD Observing and Reduction Techniques. ASP Conference Series 23, vol. 23, p. 90. Astronomical Society of the Pacific, San Francisco (1992)
16. Daubechies, I., Defrise, M., De Mol, C.: An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. Commun. Pure Appl. Math. **57**, 1413–1541 (2004)
17. Davoust, E., Pence, W.D.: Detailed bibliography on the surface photometry of galaxies. Astron. Astrophys. Suppl. Ser. **49**, 631–661 (1982)
18. Debray, B., Llebaria, A., Dubout-Crillon, R., Petit, M.: CAPELLA: software for stellar photometry in dense fields with an irregular background. Astron. Astrophys. **281**, 613–635 (1994)

19. Dempster, A., Laird, N., Rubin, D.: Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. B **39**(1), 1–38 (1977)
20. Desiderà G., Carbillet, M.: Strehl-constrained iterative blind deconvolution for post-adaptive-optics data. Astron. Astrophys. **507**(3), 1759–1762 (2009)
21. Djorgovski, S.: Modelling of seeing effects in extragalactic astronomy and cosmology. J. Astrophys. Astron. **4**, 271–288 (1983)
22. Dupé, F.-X., Fadili, M.J., Starck, J.-L.: A proximal iteration for deconvolving Poisson noisy images using sparse representations. IEEE Trans. Image Process. **18**(2), 310–321 (2009)
23. Engl, H. W., Hanke, M., Neubauer, A.: Regularization of Inverse Problems. Volume 375 of Mathematics and Its Applications. Kuwer Academic, Dordrecht/Boston (1996)
24. Figueiredo, M.A., Nowak, R.: An EM algorithm for wavelet-based image restoration. IEEE Trans. Image Process. **12**(8), 906–916 (2003)
25. Geman, S., Geman, D.: Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. IEEE Trans. Pattern Anal. Mach. Intell. **6**, 721–741 (1984)
26. Green, P.J.-F.: Bayesian reconstructions from emission tomography data using a modified EM algorithm. IEEE Trans. Med. Imaging **9**, 84–93 (1990)
27. Holmes, T.J.: Blind deconvolution of quantum-limited incoherent imagery: maximum-likelihood approach. J. Opt. Soc. Am. **A-9**, 1052–1061 (1992)
28. Irwin, M.J.: Automatic analysis of crowded fields. Mon. Not. R. Astron. Soc. **214**, 575–604 (1985)
29. Kron, R.G.: Photometry of a complete sample of faint galaxies. Astrophys. J. Suppl. Ser. **43**, 305–325 (1980)
30. Kurtz, M.J.: Classification methods: an introductory survey. In: Statistical Methods in Astronomy. European Space Agency Special Publication 201, pp. 47–58. ESA Scientific & Technical Publications Branch, Noordwijk (1983)
31. Lantéri, H., Roche, M., Aime, C.: Penalized maximum likelihood image restoration with positivity constraints: multiplicative algorithms. Inverse Probl. **18**, 1397–1419 (2002)
32. Lee, D.D., Seung, H.S.: Algorithms for non-negative matrix factorization. Adv. Neural Inf. Process. **13**, 556–562 (2001)
33. Lefèvre, O., Bijaoui, A., Mathez, G., Picat, J.P., Lelièvre, G.: Electronographic BV photometry of three distant clusters of galaxies. Astron. Astrophys. **154**, 92–99 (1986)
34. Lucy, L.B.: An iteration technique for the rectification of observed distributions. Astron. J. **79**, 745–754 (1974)
35. Maddox, S.J., Efstathiou, G., Sutherland, W.J.: The APM galaxy survey – part two – photometric corrections. Mon. Not. R. Astron. Soc. **246**, 433 (1990)
36. Mallat, S.: A Wavelet Tour of Signal Processing, The Sparse Way, 3rd edn. Academic, Boston (2008)
37. Mallat, S., Zhang, Z.: Matching pursuits with time-frequency dictionaries. IEEE Trans. Signal Process. **41**(12), 3397–3415 (1993)
38. Moffat, A.F.J.: A theoretical investigation of focal stellar images in the photographic emulsion and application to photographic photometry. Astron. Astrophys. **3**, 455–461 (1969)
39. Molina, R., Ripley, B.D., Molina, A., Moreno, F., Ortiz, J.L.: Bayesian deconvolution with prior knowledge of object location – applications to ground-based planetary images. Astrophys. J. **104**, 1662–1668 (1992)
40. Murtagh, F., Starck, J.-L., Bijaoui, A.: Image restoration with noise suppression using a multiresolution support. Astron. Astrophys. Suppl. Ser. **112**, 179–189 (1995)
41. Natterer, F., Wûbbeling, F.: Mathematical Methods in Image Reconstruction. SIAM, Philadelphia (2001)
42. Naylor, T.: An optimal extraction algorithm for imaging photometry. Mon. Not. R. Astron. Soc. **296**, 339–346 (1998)
43. Okamura, S.: Global structure of Virgo cluster galaxies. In: ESO Workshop On The Virgo Cluster of Galaxies, Garching, pp. 201–215 (1985)
44. Pence, W.D., Davoust, E.: Supplement to the detailed bibliography on the surface photometry of galaxies. Astron. Astrophys. Suppl. Ser. **60**, 517–526 (1985)

45. Pierre, M., Valtchanov, I., Altieri, B., Andreon, S., Bolzonella, M., Bremer, M., Disseau, L., Dos Santos, S., Gandhi, P., Jean, C., Pacaud, F., Read, A., Refregier, A., Willis, J., Adami, C., Alloin, D., Birkinshaw, M., Chiappetti, L., Cohen, A., Detal, A., Duc, P., Gosset, E., Hjorth, J., Jones, L., LeFevre, O., Lonsdale, C., Maccagni, D., Mazure, A., McBreen, B., McCracken, H., Mellier, Y., Ponman, T., Quintana, H., Rottgering, H., Smette, A., Surdej, J., Starck, J., Vigroux, L., White, S.: The XMM-LSS survey. Survey design and first results. J. Cosmol. Astro-Part. Phys. **9**, JCAP09(2004)011 (2004)
46. Prato, M., Cavicchioli, R., Zanni, L., Boccacci, P., Bertero, M.: Efficient deconvolution methods for astronomical imaging: algorithms and IDL-GPU codes. Astron. Astrophys. **539**, A133 (2012)
47. Prato, M., La Camera, A., Bonettini, S., Bertero, M.: A convergent blind deconvolution method for post-adaptive-optics astronomical imaging. Inverse Probl. **29**(6), 065017 (2013)
48. Richardson, W.H.: Bayesian-based iterative method of image restoration. J. Opt. Soc. Am. **62**, 55–59 (1972)
49. Shepp, L.A., Vardi, Y.: Maximum likelihood reconstruction for emission tomography. IEEE Trans. Med. Imaging **MI-2**, 113–122 (1982)
50. Starck, J.-L., Aussel, H., Elbaz, D., Fadda, D., Cesarsky, C.: Faint source detection in ISOCAM images. Astron. Astrophys. Suppl. Ser. **138**, 365–379 (1999)
51. Starck, J.-L., Bijaoui, A., Murtagh, F.: Multiresolution support applied to image filtering and deconvolution. CVGIP: Graph. Models Image Process. **57**, 420–431 (1995)
52. Starck, J.-L., Elad, M., Donoho, D.L.: Redundant multiscale transforms and their application for morphological component analysis. Adv. Imaging Electron Phys. **132**, 287–348 (2004)
53. Starck, J.-L., Fadili, J., Murtagh, F.: The undecimated wavelet decomposition and its reconstruction. IEEE Trans. Image Process. **16**, 297–309 (2007)
54. Starck, J.-L., Murtagh, F.: Image restoration with noise suppression using the wavelet transform. Astron. Astrophys. **288**, 343–348 (1994)
55. Starck, J.-L., Murtagh, F.: Automatic noise estimation from the multiresolution support. Publ. Astron. Soc. Pac. **110**, 193–199 (1998)
56. Starck, J.-L., Murtagh, F.: Astronomical Image and Data Analysis. Springer, Berlin (2002).
57. Starck, J.-L., Murtagh, F.: Astronomical Image and Data Analysis, 2nd edn. Springer, Berlin (2006)
58. Starck, J.-L., Murtagh, F., Bijaoui, A.: Image Processing and Data Analysis: The Multiscale Approach. Cambridge University Press, Cambridge/New York (1998)
59. Starck, J.-L., Pierre, M.: Structure detection in low intensity X-ray images. Astron. Astrophys. Suppl. Ser. **128**, 397–407 (1998).
60. Starck, J.-L., Siebenmorgen, R., Gredel, R.: Spectral analysis by the wavelet transform. Astrophys. J. **482**, 1011–1020 (1997)
61. Takase, B., Kodaira, K., Okamura, S.: An Atlas of Selected Galaxies. University of Tokyo Press, Tokyo (1984)
62. Thonnat, M.: INRIA Rapport de Recherche, Centre Sophia Antipolis, No. 387 (1985). Automatic morphological description of galaxies and classification by an expert system
63. Tikhonov, A.N., Goncharski, A.V., Stepanov, V.V., Kochikov, I.V.: Ill-posed image processing problems. Sov. Phys. – Dokl. **32**, 456–458 (1987)
64. Watanabe, M., Kodaira, K., Okamura, S.: Digital surface photometry of galaxies toward a quantitative classification. I. 20 galaxies in the Virgo cluster. Astron. Astrophys. Suppl. Ser. **50**, 1–22 (1982)
65. Zanella, R., Boccacci, P., Zanni, L., Bertero, M.: Efficient gradient projection methods for edge-preserving removal of Poisson noise. Inverse Probl. **25**, 045010 (2009)
66. Zhang, B., Fadili, M.J., Starck, J.-L.: Wavelets, ridgelets and curvelets for Poisson noise removal. IEEE Trans. Image Process. **17**(7), 1093–1108 (2008)

Differential Methods for Multi-dimensional Visual Data Analysis

Werner Benger, René Heinzl, Dietmar Hildenbrand, Tino Weinkauf, Holger Theisel, and David Tschumperlé

Contents

1 Introduction 2100
2 Modeling Data via Fiber Bundles 2102
Differential Geometry: Manifolds, Tangential Spaces, and Vector Spaces 2103
Topology: Discretized Manifolds 2111
Ontological Scheme and Seven-Level Hierarchy 2113
3 Differential Forms and Topology 2119
Differential Forms 2119
Homology and Cohomology 2128

W. Benger (✉)
Airborne Hydromapping Software GmbH, Innsbruck, Austria

Institute for Astro- and Particle Physics, University of Innsbruck, Innsbruck, Austria

Center for Computation and Technology, Lousiana State University, Baton Rouge, LA, USA
e-mail: werner@cct.lsu.edu

R. Heinzl
Shenteq s.r.o, Bratislava, Slovak Republic
e-mail: heinzl@shenteq.com

D. Hildenbrand
University of Technology Darmstadt, Darmstadt, Germany

T. Weinkauf
Feature-Based Data Analysis for Computer Graphics and Visualization, Max Planck Institute for Informatics, Saarbrücken, Germany
e-mail: weinkauf@mpi-inf.mpg.de

Holger Theisel
Institut fur Simulation und Graphik AG Visual Computing, Magdeburg, Germany
e-mail: theisel@isg.cs.uni-magdeburg.de

David Tschumperlé
GREYC (UMR-CNRS 6072), CAEN Cedex, France
e-mail: david.tschumperle@greyc.ensicaen.fr

© Springer Science+Business Media New York 2015

O. Scherzer (ed.), *Handbook of Mathematical Methods in Imaging*,
DOI 10.1007/978-1-4939-0790-8_35

Topology 2130
4 Geometric Algebra Computing 2132
Benefits of Geometric Algebra 2134
Conformal Geometric Algebra 2136
Computational Efficiency of Geometric Algebra Using Gaalop 2139
5 Feature-Based Vector Field Visualization 2141
Characteristic Curves of Vector Fields 2141
Derived Measures of Vector Fields 2143
Topology of Vector Fields 2145
6 Anisotropic Diffusion PDEs for Image Regularization and Visualization 2149
Regularization PDEs: A Review 2149
Applications 2154
7 Conclusion 2157
Cross-References 2159
References 2159

Abstract

Images in scientific visualization are the end product of data processing. Starting from higher-dimensional data sets such as scalar, vector, and tensor fields given on 2D, 3D, and 4D domains, the objective is to reduce this complexity to two-dimensional images comprehensible to the human visual system. Various mathematical fields such as in particular differential geometry, topology (theory of discretized manifolds), differential topology, linear algebra, Geometric Algebra, vector field and tensor analysis, and partial differential equations contribute to the data filtering and transformation algorithms used in scientific visualization. The application of differential methods is core to all these fields. The following chapter will provide examples from current research on the application of these mathematical domains to scientific visualization. Ultimately the use of these methods allows for a systematic approach for image generation resulting from the analysis of multidimensional datasets.

1 Introduction

> Scientists need an alternative to numbers. The use of images is a technical reality nowadays and tomorrow it will be an essential requisite for knowledge. The ability of scientists to visualize calculations and complex simulations is absolutely essential to ensure the integrity of analyses, to promote scrutiny in depth and to communicate the result of such scrutiny to others... The purpose of scientific calculation is looking, not enumerating. It is estimated that 50 % of the brain's neurons are associated with vision. Visualization in a scientific calculation is aimed at putting this neurological machinery to work [56].

Since this visionary quote from an article in 1987, scientific visualization, benefiting from the affordable graphics hardware driven by the computer gaming industry, has grown rapidly. Beyond academic research interests it has become also a consumer market with practical applicability in industry and medicine. Still there are yet many gaps that are left open due to the unequal evolution velocities in different fields. Once, there is the human mind that is not able to keep up with the deluge of visual

information which can be produced with modern technology. Many scientists still prefer looking at numbers instead of utilizing modern display technology. At the same time, data can be produced by modern supercomputers that is far beyond the ability of even high-end graphics engines to be processed. Data sets originating from numerical simulations of physical processes will usually be three-dimensional or four-dimensional, with images just the final result of the process of scientific visualization. In this context images are the means to analyze data set of higher dimensions.

Reducing numerical data sets to images is known as the concept of the *visualization pipeline*. In its simplest form it consists of a data source (n-dimensional), a data filter (an algorithmic operation), and a data sink (an image). Data filters need to understand the structure and meaning of the multidimensional input data and to operate efficiently on them. This involves various mathematical fields such as in particular differential geometry, topology (theory of discretized manifolds), differential topology, linear algebra, Geometric Algebra, vector field and tensor analysis, and partial differential equations. Within a scientific visualization process, all these mathematical fields will work together, with more or less weighting. We subsume this set of mathematical domains as "differential methods" in this chapter as the concept of differentiation is fundamental to their approach of data analysis. The following sections will demonstrate the application of the respective mathematical fields to visual analysis by virtue of examples of ongoing research.

In Sect. 2 we discuss the general issue of how to lay out data to model the structure of space and time, as we know it from mathematics as foundation for further operations. Frequently visualization algorithms are implemented ad hoc, given the problem, inventing the solution with highest performance. This allegedly reasonable approach comes with an unfortunate downside: incompatibility among independently developed solutions, which impacts data exchange and interfacing complementary implementations. However, when keeping a common data model in mind right from the earliest steps of conceiving some algorithm, interoperability can be achieved at no cost with same performance as solitary solutions.

Given a solid foundation for data structures, Sect. 3 demonstrates how to formulate differential operators using the concepts of chains, cochains, homology, and cohomology. Since in computer graphics and visualization we have to deal with discretized spaces, we arrive in the mathematical field of topology, as an essential descriptive tool for meshes and all nontrivial grid structures.

When considering mathematics as a language unifying computer science, we need to even more think about a common denominator within mathematics itself. Geometric Algebra is a relatively new – or, rather, rediscovered – branch of mathematics that is very promising. It is extraordinarily visually intuitive while covering the abstractions of Clifford algebra as used in quantum mechanics equally well as the formulations of curved space in general relativity. However, even independent of such physics-oriented applications, Geometric Algebra has found its merits within computer graphics itself. Section 4 will talk about the elegant usage of five-dimensional projective conformal Geometric Algebra to handle primitives in

computer graphics and eventually implement the ray-tracing algorithm with a few, well-defined algebraic operations.

The general goal of visualization is to give insight into large and complex data sets. Due to the sheer size of the data sets alone, it is favorable if not necessary to automate at least parts of the analysis. A way to achieve this is by extracting features. Features can either be certain quantities derived from a data set or a mathematically well-defined, geometric object (point, line, surface, ...) with its definition and interpretation depending on the underlying application, but usually it represents important structures (e.g., vortex, stagnation point) or changes to such structures (events, bifurcations). A feature-based visualization aims at the reduction of information to guide a user to the most interesting parts of a data set. In Sect. 5 we describe some important approaches to feature-based visualization of vector fields. These include investigation of derived quantities such as vortices (section "Derived Measures of Vector Fields") and the topology of vector fields (section "Topology of Vector Fields"). These approaches have become a standard tool for the analysis of vector fields.

Finally, in Sect. 6 we explore the capabilities of partial differential equations for the filtering and regularization of image data sets. Applications are enhancing image quality by reducing noise or similar artifacts, as well as the visualization of vector and tensor fields.

2 Modeling Data via Fiber Bundles

Purely numerical algorithms in C++ can be abstracted from concrete data structures using programming techniques such as generic programming [79]. However, generic algorithms still need to make certain assumptions about the data they operate on. The question remains what these *concepts* are that describe "data": what properties should be expected by some algorithm from any kind of data provided for scientific visualization? Moreover, consistency among concepts shared by independent algorithms is also required to achieve *interoperability* among algorithms and eventually (independently developed) applications. While any particular problem can be addressed by some particular solution, a common concept allows to build a *framework* instead of just a collection of *tools*. Tools are what an end user needs to solve a particular problem with a known solution. However, when a problem is not yet clearly defined and a solution unknown, then a framework is required that allows exploration of various approaches and eventually adaption toward a specific direction that does not exist a priori.

The concept of how to lay out data to perform visualization operations in a common framework constitutes a *data model* for visualization. Many visualization applications are to a greater or lesser extent a collection of tools, even when bundled together within the same software library or binary. Consequently, interoperability between different applications and their corresponding file formats is hard or impossible. Only very few implementations adhere to the vision of a common data model across the various data types for visualization. The idea of a common data

model is frequently undervalued or even disregarded as being impossible. However, as D. Butler said, "The proper abstractions for scientific data are known. We just have to use them" [16].

D. Butler was following the mathematical concepts of fiber bundles [16], or more specific, vector bundles [15], to model data. The IBM Data Explorer, one of the earliest visualization applications, now open source and known as "OpenDX (http://www.opendx.org)," implemented this concept successfully [76]. These ideas have been revived and expanded by [7] leading to a hierarchical data structure consisting of a noncyclic graph in seven levels. It can be seen as largely keyword-free, hierarchical version of the OpenDX model, seeking to cast the information and relationships provided in original model into a grouping structure. This data model will be reviewed in the following, together with its mathematical background. Section "Differential Geometry: Manifolds, Tangential Spaces, and Vector Spaces" will review the basic mathematical structures that are used to describe space and time. Section "Topology: Discretized Manifolds" will introduce the mathematical formulation of discretized space. Based on this background, section "Ontological Scheme and Seven-Level Hierarchy" will present a scheme that is able to cover the described mathematical structures.

Differential Geometry: Manifolds, Tangential Spaces, and Vector Spaces

Space and time in physics is modeled via the concept of a differentiable manifold. As scientific visualization deals with data given in space and time, following these concepts is reasonable. In short, a manifold is a topological space that is locally homeomorphic to $\mathbb{R}^n$. However, not all data occurring in scientific visualization are manifolds. The more general case of topological spaces will be discussed in sections "Topology: Discretized Manifolds" and "Topology."

A vector space over a field F(such as $\mathbb{R}$) is a set V together with two binary operations *vector addition* $+: V \times V \to V$ and *scalar multiplication* $\circ : F \times V \to V$. The mathematical concept of a *vector* is defined as an element $v \in V$. A vector space is closed under the operations $+$ and $\circ$, i.e., for all elements $u, v \in V$ and all elements $\lambda \in F$ there is $u + v \in V$ and $\lambda \circ u \in V$ (vector space axioms). The vector space axioms allow computing the differences of vectors and therefore defining the derivative of a vector-valued function $v(s) : \mathbb{R} \to V$ as

$$\frac{d}{ds}v(s) := \lim_{ds \to 0} \frac{v(s + ds) - v(s)}{ds} \tag{1}$$

A manifold in general is *not* a vector space. However, a differentiable manifold M allows to define a tangential space $T_P(M)$ at each point P which has vector space properties.

Tangential Vectors

In differential geometry, a tangential vector on a manifold M is the operator $\frac{d}{ds}$ that computes the derivative along a curve $q(s) : \mathbb{R} \to M$ for an arbitrary scalar-valued function $f : M \to \mathbb{R}$:

$$\frac{d}{ds} f \Big|_{q(s)} := \frac{df(q(s))}{ds} \tag{2}$$

Tangential vectors fulfill the vector space axioms and can therefore be expressed as linear combinations of derivatives along the n coordinate functions $x^\mu : M \to \mathbb{R}$ with $\mu = 0 \ldots n-1$, which define a basis of the tangential space $T_{q(s)}(M)$ on the n-dimensional manifold M at each point $q(s) \in M$:

$$\frac{d}{ds} f = \sum\nolimits_{\mu=1}^{n-1} \frac{dx^\mu(q(s))}{ds} \frac{\partial}{\partial x^\mu} f =: \sum\nolimits_{\mu=1}^{n-1} \dot{q}^\mu \partial_\mu f \tag{3}$$

where $\{q\}^\mu$ are the components of the tangential vector $\frac{d}{ds}$ in the chart $\{x^\mu\}$ and $\{\partial_\mu\}$ are the basis vectors of the tangential space in this chart. In the following text the Einstein sum convention is used, which assumes implicit summation over indices occurring on the same side of an equation. Often tangential vectors are used synonymous with the term "vectors" in computer graphics when a direction vector from point A to point B is meant. A tangential vector on an n-dimensional manifold is represented by n numbers in a chart.

Covectors

The set of operations df: $T(M) \to \mathbb{R}$ that map tangential vectors $v \in T(M)$ to a scalar value $v(f)$ for any function $f : M \to \mathbb{R}$ defines another vector space which is dual to the tangential vectors. Its elements are called *covectors*:

$$< df, v >= df(v) := v(f) = v^\mu \partial_\mu f = v^\mu \frac{\partial f}{\partial x^\mu} \tag{4}$$

Covectors fulfill the vector space axioms and can be written as linear combination of covector basis functions dx^μ:

$$df =: \frac{\partial f}{\partial x^\mu} dx^\mu \tag{5}$$

whereby the dual basis vectors fulfill the duality relation

$$< dx^\nu, \partial_\mu >= \begin{cases} \mu = \nu : & 1 \\ \mu \neq \nu : & 0 \end{cases} \tag{6}$$

The space of covectors is called the cotangential space $T_P{}^*(M)$. A covector on an n-dimensional manifold is represented by n numbers in a chart, same as a tangential vector. However, covectors transform inverse to tangential vectors when changing

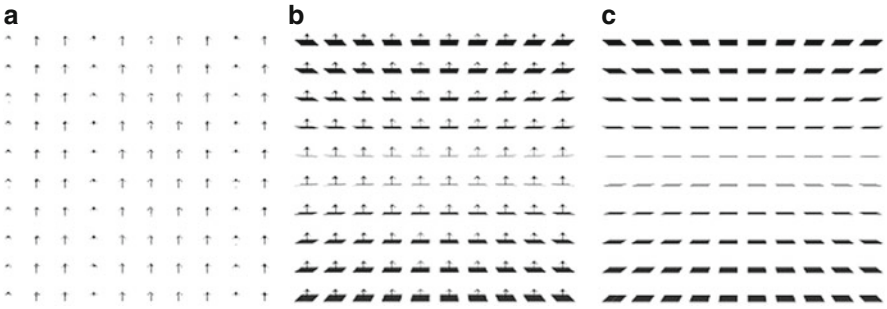


Fig. 1 The (trivial) constant vector field along the z-axis viewed as vector field ∂_z and as covector field dz. (**a**) Vector field ∂_z. (**b**) Duality relationship among ∂_z and dz. (**c**) Co-vector field dz

coordinate systems, as is directly obvious from Eq. (6) in the one-dimensional case: as $< dx^0, \partial_0 >= 1$ must be sustained under coordinate transformation, dx^0 must shrink by the same amount as ∂_0 grows when another coordinate scale is used to represent these vectors. In higher dimensions this is expressed by an inverse transformation matrix.

In Euclidean three-dimensional space, a plane is equivalently described by a "normal vector," which is orthogonal to the plane. While "normal vectors" are frequently symbolized by an arrow, similar to tangential vectors, they are not the same, rather they are dual to tangential vectors. It is more appropriate to visually symbolize them as a plane. This visual is also supported by (5), which can be interpreted as the total differential of a function f: a covector describes the change of a function f along a direction as specified by a tangential vector $\vec{v}$. A covector V can thus be visually imagined as a sequence of coplanar (locally flat) planes at distances given by the magnitude of the covector that count the number of planes which are crossed by a vector $\vec{w}$. This number is $V(w)$. For instance, for the Cartesian coordinate function x, the covector dx "measures" the "crossing rate" of a vector w in the direction along the coordinate line x; see Figs. 1 and 2. On an n-dimensional manifold a covector is correspondingly symbolized by a $(n-1)$-dimensional subspace.

Tensors

A *tensor* T_n^m of rank $n \times m$ is a multi-linear map of n vectors and m covectors to a scalar

$$T_n^m : T(M) \times \ldots T(M)_n \times T^*(M) \times \ldots T^*(M)_m \rightarrow \mathbb{R} \tag{7}$$

Tensors are elements of a vector space themselves and form the tensor algebra. They are represented relative to a coordinate system by a set of k^{n+m} numbers for a k-dimensional manifold. Tensors of rank 2 may be represented using matrix notation. Tensors of type $T_1{}^0$ are equivalent to covectors and called co-variant; in

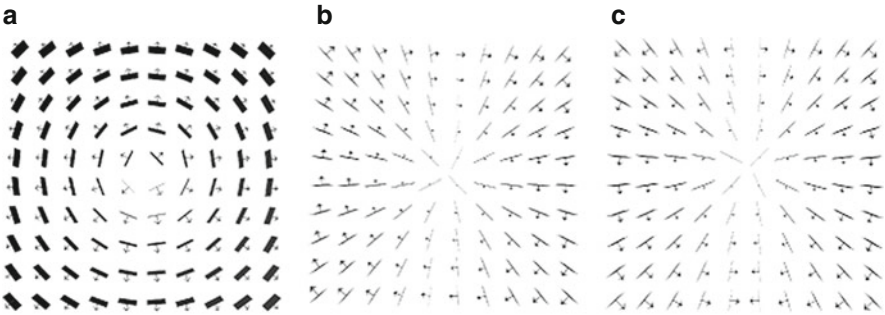


Fig. 2 The basis vector and covector fields induced by the polar coordinates $\{r, \vartheta, \phi\}$. (**a**) Radial field dr ∂_r. (**b**) Azimuthal field dϕ ∂_ϕ view of the equatorial plane (z-axis towards eye). (**c**) Altitudal field dθ ∂_θ slice along the z-axis

matrix notation (relative to a chart) they correspond to rows. Tensors of type $T_0{}^1$ are equivalent to a tangential vector and are called contra-variant, corresponding to columns in matrix notation. The duality relationship between vectors and covectors then corresponds to the matrix multiplication of a $1 \times n$ row with a $n \times 1$ column, yielding a single number

$$< a, b > = < a^\mu \partial_\mu, b_\mu dx^\mu > \equiv \ (a^0 a^1 \dots a^n) \begin{pmatrix} b^0 \\ b^1 \\ \dots \\ b^n \end{pmatrix} \tag{8}$$

By virtue of the duality relationship (6), the contraction of lower and upper indices is defined as the *interior product* ι of tensors, which reduces the dimensionality of the tensor:

$$\iota : T_n^m \times T_k^l \to T_{n-1}^{m-k} : u, v \mapsto \iota_u v \tag{9}$$

The interior product can be understood (visually) as a generalization of some "projection" of a tensor onto another one.

Of special importance are symmetric tensors of rank two $g \in T_2^0$ with $g : T(M) \times T(M) \to \mathbb{R} : u, v \mapsto g(u, v), g(u, v) = g(v, u)$, as they can be used to define a *metric* or *inner product* on the tangential vectors. Its inverse, defined by operating on the covectors, is called the co-metric. A metric, same as the co-metric, is represented as a symmetric $n \times n$ matrix in a chart for an n-dimensional manifold.

Given a metric tensor, one can define equivalence relationships between tangential vectors and covectors, which allow to map one into each other. These maps are called the "musical isomorphisms," $\flat$ and $\sharp$, as they raise or lower an index in the coordinate representation:

$$\flat: \quad T(M) \to T^*(M): \quad v^\mu \partial_\mu \mapsto v^\mu g_{\mu\nu} dx^\nu \tag{10}$$

$$\sharp: \quad T^*(M) \to T(M): \quad V_\mu dx^\mu \mapsto V_\mu g^{\mu\nu} \partial_\nu \tag{11}$$

As an example application, the "gradient" of a scalar function is given by $\nabla f = \sharp df$ using this notation. In Euclidean space, the metric is represented by the identity matrix and the components of vectors are identical to the components of covectors. As computer graphics usually is considered in Euclidean space, this justifies the usual negligence of distinction among vectors and covectors; consequently graphics software only knows about one type of vectors which is uniquely identified by its number of components. However, when dealing with coordinate transformations or curvilinear mesh types, distinguishing between tangential vectors and covectors is unavoidable. Treating them both as the same type within a computer program leads to confusions and is not safe.

Exterior Product

The *exterior product* $\wedge : V \times V \to \Lambda(V)$ is an algebraic construction generating vector space elements of higher dimensions from elements of a vector space V. The new vector space is denoted $\Lambda(V)$. It is alternating, fulfilling the property $v \wedge u = -u \wedge v \ \forall u, v \in V$ (which results in $v \wedge v = 0 \ \forall v \in V$). The exterior product defines an algebra on its elements, the exterior algebra (or Grassmann algebra). It is a sub-algebra of the tensor algebra consisting of the antisymmetric tensors. The exterior algebra is defined intrinsically by the vector space and does not require a metric. For a given n – *dimensional* vector space V, there can at most be nth power of an exterior product, consisting of n different basis vectors. The $(n + 1)$th power must vanish, because at least one basis vector would occur twice, and there is exactly one basis vector in $\Lambda^n(V)$.

Elements $v \in \Lambda^k(V)$ are called k-vectors, whereby two-vectors are also called bi-vectors and three-vectors tri-vectors. The number of components of a k-vector of an n-dimensional vector space is given by the binomial coefficient $\{n\}\{k\}$. For $n = 2$ there are two one-vectors and one bi-vector, for $n = 3$ there are three one-vectors, three bi-vectors, and one tri-vector. These relationships are depicted by the Pascal's triangle, with the row representing the dimensionality of the underlying base space and the column the vector type:

$$\begin{array}{ccccccc} & & & 1 & & & \\ & & 1 & & 1 & & \\ & 1 & & 2 & & 1 & \\ 1 & & 3 & & 3 & & 1 \\ 1 & & 4 & & 6 & & 4 & & 1 \end{array} \tag{12}$$

As can be easily read off, for a four-dimensional vector space, there will be four one-vectors, six bi-vectors, four tri-vectors, and one four-vector. The n-vector of

an n-dimensional vector space is also called a *pseudoscalar*, the $(n-1)$ vector a *pseudo-vector*.

Visualizing Exterior Products

An exterior algebra is defined on both the tangential vectors and covectors on a manifold. A bi-vector v formed from tangential vectors is written in chart as

$$v = v^{\mu\nu} \partial_\mu \wedge \partial_\nu \tag{13}$$

and a bi-covector U formed from covectors is written in chart as

$$U = U_{\mu\nu} dx^\mu \wedge dx^\nu \tag{14}$$

They both have $\{n\}\{2\}$ independent components, due to $v^{\mu\nu} = -v^{\nu\mu}$ and $U_{\mu\nu} = -U_{\nu\mu}$ (three components in 3D, six components in 4D). A bi-tangential vector (13) can be understood visually as an (oriented, i.e., signed) plane that is spun by the two defining tangential vectors, independently of the dimensionality of the underlying base space. A bi-covector (14) corresponds to the subspace of an n-dimensional hyperspace where a plane is "cut out." In three dimensions these visualizations overlap: both a bi-tangential vector and a covector correspond to a plane, and both a tangential vector and a bi-covector correspond to one-dimensional direction ("arrow"). In four dimensions, these visuals are more distinct but still overlap: a covector corresponds to a three-dimensional volume, but a bi-tangential vector is represented by a plane same as a bi-covector, since cutting out a 2D plane from four-dimensional space yields a 2D plane again. Only in higher dimensions these symbolic representations become unique. However, both a co-vector and a pseudo-vector will always correspond to (i.e., appear as) an $(n-1)$-dimensional hyperspace.

$$V_\mu dx^\mu \Longleftrightarrow v_{\alpha_0 \alpha_1 \ldots \alpha_{n-1}} \partial_{\alpha_0} \wedge \partial_{\alpha_1} \wedge \ldots \partial_{\alpha_{n-1}} \tag{15}$$

$$v^\mu \partial_\mu \Longleftrightarrow V_{\alpha_0 \alpha_1 \ldots \alpha_{n-1}} dx^{\alpha_0} \wedge dx^{\alpha_1} \wedge \ldots dx^{\alpha_{n-1}} \tag{16}$$

A tangential vector – lhs of (16) – can be understood as one specific direction. Equivalently, it can be seen as "cutting off" all but one $(n-1)$-dimensional hyperspaces from the full n-dimensional space. This equivalence is expressed via the interior product of a tangential vector v with a pseudo-co-scalar Ω yielding a pseudo-covector V (17). Similarly, the interior product of a pseudo-vector with a pseudo-co-scalar yields a tangential vector (17):

$$\iota_\Omega : T(M) \to (T^*)^{(n-1)}(M) : V \mapsto \iota_\Omega v \tag{17}$$

$$\iota_\Omega : T^{(n-1)}(M) \to T^*(M) : V \mapsto \iota_\Omega v \tag{18}$$

Pseudoscalars and pseudo-co-scalars will always be scalar multiples of the basis vectors $\partial_{\alpha 0} \wedge \partial_{\alpha 1} \wedge \ldots \partial_{\alpha n}$ and $dx_0^\alpha \wedge dx_1^\alpha \wedge \ldots dx_n^\alpha$. However, when inversing a coordinate $x^\mu \to -x^\mu$, they flip sign, whereas a "true" scalar does not. An example known from Euclidean vector algebra is the allegedly scalar value constructed from the dot and cross product of three vectors $V(u, v, w) = u \cdot (v \times w)$ which is the negative of when its arguments are flipped:

$$V(u, v, w) = -V(-u, -v, -w) = -u \cdot (-v \times -w) \tag{19}$$

which is actually more obvious when (19) is written as exterior product:

$$V(u, v, w) = u \wedge v \wedge w = V \partial_0 \wedge \partial_1 \wedge \partial_2 \tag{20}$$

The result (20) actually describes the multiple of a volume element span by the basis tangential vectors ∂_μ – any volume must be a scalar multiple of this basis volume element but can flip sign if another convention on the basis vectors is used. This convention depends on the choice of a right-handed versus left-handed coordinate system and is expressed by the orientation tensor $\Omega = \pm \partial_0 \wedge \partial_1 \wedge \partial_2$. In computer graphics, both left-handed and right-handed coordinate systems occur, which may lead to lots of confusions.

By combining (18) and (11) – requiring a metric – we get a map from pseudo-vectors to vectors and reverse. This map is known as the *Hodge star operator* "*":

$$_* : T^{(n-1)}(M) \to T(M) : V \vdash \to \sharp \iota_\Omega V \tag{21}$$

The same operation can be applied to the covectors accordingly and generalized to all vector elements of the exterior algebra on a vector space, establishing a correspondence between k – *vectors* and $n - k$-vectors. The Hodge star operator allows to identify vectors and pseudo-vectors, similar to how a metric allows to identify vectors and covectors. The Hodge star operator requires a metric and an orientation Ω.

A prominent application in physics using the hodge star operator are the Maxwell equations, which, when written based on the four-dimensional potential $A = V_0 dx^0 + A_k dx^k$ (V_0 the electrostatic, A_k the magnetic vector potential), take the form

$$d_* dA = J \tag{22}$$

with J the electric current and magnetic flow, which is zero in vacuum. The combination $d * d$ is equivalent to the Laplace operator "$\square$," which indicates that (22) describes electromagnetic waves in vacuum.

Geometric Algebra

Geometric Algebra is motivated by the intention to find a closed algebra on a vector space with respect to multiplication, which includes existence of an inverse

operation. There is no concept of dividing vectors in "standard" vector algebra. Neither the inner or outer product has provided vectors of the same dimensionality as their arguments, so they do not provide a closed algebra on the vector space.

Geometric Algebra postulates a product on elements of a vector space $u, v, w \in \mathcal{V}$ that is associative $(uv)w = u(vw)$, left distributive $u(v + w) = uv + uw$, and right distributive $(u + v)w = uw + vw$ and reduces to the inner product as defined by the metric $v^2 = g(v, v)$. It can be shown that the sum of the outer product and the inner product fulfill these requirements; this defines the *geometric product* as the sum of both:

$$uv := u \wedge v + u \cdot v \tag{23}$$

Since $u \wedge v$ and $u \cdot v$ are of different dimensionalities ($\{n\}\{\{2\}$ and $\{n\}\{\{0\}$, respectively), the result must be in a higher-dimensional vector space of dimensionality $\{n\}\{\{2\} + \{n\}\{\{0\}$. This space is formed by the linear combination of k-vectors; its elements are called *multivectors*. Its dimensionality is $\sum_{k=0}^{n-1} \binom{n}{k} \equiv 2^n$.

For instance, in two dimensions, the dimension of the space of multivectors is $2^2 = 4$. A multivector V, constructed from tangential vectors on a two-dimensional manifold, is written as

$$V = V^0 + V^1 \partial_0 + V^2 \partial_1 + V^3 \partial_0 \wedge \partial_1 \tag{24}$$

with V^μ the four components of the multivector in a chart. For a three-dimensional manifold, a multivector on its tangential space has $2^3 = 8$ components and is written as

$$\begin{aligned} V = {} & V^0 + \\ & V^1 \partial_0 + V^2 \partial_1 + V^2 \partial_2 + \\ & V^4 \partial_0 \wedge \partial_1 + V^5 \partial_1 \wedge \partial_2 + V^6 \partial_2 \wedge \partial_0 + \\ & V^7 \partial_0 \wedge \partial_1 \wedge \partial_2 \end{aligned} \tag{25}$$

with V^μ the eight components of the multivector in a chart. The components of a multivector have a direct visual interpretation, which is one of the key features of Geometric Algebra. In 3D, a multivector is the sum of a scalar value, three directions, three planes, and one volume. These basis elements span the entire space of multivectors. Geometric Algebra provides intrinsic graphical insight to the algebraic operations. Its application for computer graphics will be discussed in Sect. 4.

Vector and Fiber Bundles

The concept of a fiber bundle data model is inspired by its mathematical correspondence. In short, a fiber bundle is a topological space that looks locally like a product space $B \times F$ of a base space B and a fiber space F.

The *fibers of a function* $f\colon X \to Y$ are the pre-images or inverse images of the points $y \in Y$, i.e., the sets of all elements $x \in X$ with $f(x) = y$:

$$f^{-1}(y) = \{x \in X | f(x) = y\}$$

is a fiber of f (at the point y). A fiber can also be the empty set. The union set of all fibers of a function is called the *total space*. The definition of a fiber bundle makes use of a *projection mappr*$_1$, which is a function that maps each element of a product space to the element of the first space:

$$\begin{array}{rcl} pr_1 : X \times Y & \to & X \\ (x, y) & \longmapsto & x \end{array}$$

Let E, B be topological spaces and $f\colon E \to B$ a continuous map. (E, B, f) is called a *(fiber) bundle* if there exists a space F such that the union of fibers of a neighborhood $U_b \subset B$ of each point $b \in B$ is homeomorphic to $U_b \times F$ such that the projection pr_1 of $U_b \times F$ is U_b again:

$$\begin{array}{c} (E, B, f : E \to B) \text{ bundle} \Longleftrightarrow \exists F : \forall b \in B : \exists U_b : f^{-1}(U_b) \overset{\text{hom}}{\simeq} U_b \times F \\ \text{and} \quad pr_1(U_b \times F) = U_b \end{array}$$

E is called the *total space* E, B is called the *base space*, and $f : E \to B$ the *projection map*. The space F is called the *fiber type* of the bundle or simply the *fiber* of the bundle. In other words, the total space can be written locally as a product space of the base space with some space F. The notation $\mathcal{F}(B) = (E, B, f)$ will be used to denote a fiber bundle over the base space B. It is also said that the *space F fibers over the base space B*.

An important case is the *tangent bundle*, which is the union of all tangent spaces $T_p(M)$ on a manifold M together with the manifold $\mathcal{T}(M) := \{(p, v) : p \in M, v \in T_p(M)\}$. Every differentiable manifold possesses a tangent bundle $\mathcal{T}(M)$. The dimension of $\mathcal{T}(M)$ is twice the dimension of the underlying manifold M, its elements are points plus tangential vectors. $T_p(M)$ is the fiber of the tangent bundle over the point p.

If a fiber bundle over a space B with fiber F can be written as $B \times F$ globally, then it is called a *trivial bundle* $(B \times F, B, pr_1)$. In scientific visualization, usually only trivial bundles occur. A well-known example for a nontrivial fiber bundle is the Möbius strip.

Topology: Discretized Manifolds

For computational purposes, a topological space is modeled by a finite set of points. Such a set of points intrinsically carries a discrete topology by itself, but one usually

considers embeddings in a space that is homeomorphic to Euclidean space to define various structures describing their spatial relationships.

A subset $c \subset X$ of a Hausdorff space X is a *k-cell* if it is homeomorphic to an open k-dimensional ball in $\mathbb{R}^n$. The dimension of the cell is k. Zero-cells are called vertices, one-cells are edges, two-cells are faces or polygons, and three-cells are polyhedra – see also section "Chains." An n-cell within an n-dimensional space is just called a "cell." $(n-1)$-cells are sometimes called "facets" and $(n-2)$-cells are known as "ridges." For k-cells of arbitrary dimension, incidence and adjacency relationships are defined as follows: two cells c_1, c_2 are *incident* if $c_1 \subseteq \partial c_2$, where ∂c_2 denotes the border of the cell c_2. Two cells of the same dimension can never be incident because $\dim(c_1) \neq \dim(c_2)$ for two incident cells c_1, c_2. c_1 is a *side* of c_2 if $\dim(c_1) < \dim(c_2)$, which may be written as $c_1 < c_2$. The special case $\dim(c_1) = \dim(c_2) - 1$ may be denoted by $c_1 \prec c_2$. Two k -cells c_1, c_2 with $k > 0$ are called *adjacent* if they have a common side, i.e.,

$$\text{cell } c_1, c_2 \quad \textbf{adjacent} \iff \exists \quad \text{cell } f : \; f < c_1, f < c_2$$

For $k = 0$, two zero-cells (i.e., vertices) v_1, v_2 are said to be adjacent if there exists a one-cell (edge) e which contains both, i.e., $v_1 < e$ and $v_2 < e$. Incidence relationships form an incidence graph. A path within an incidence graph is a cell tuple: a *cell-tuple* $\mathcal{C}$ within an n-dimensional Hausdorff space is an ordered sequence of k-cells $(c_n, c_{n-1}, \ldots, c_1, c_0)$ of decreasing dimensions such that $\forall 0 < i \leq n : c_{i-1} \prec c_i$. These relationships allow to determine topological neighborhoods: adjacent cells are called *neighbors*. The set of all $k+1$ cells which are incident to a k-cell forms a neighborhood of the k-cell. The cells of a Hausdorff space X constitute a topological base, leading to the following definition: a ("closure-finite, weak-topology") *CW-complex* $\mathcal{C}$, also called a *decomposition* of a Hausdorff space X, is a hierarchical system of spaces $X^{(-1)} \subseteq X^{(0)} \subseteq X^{(1)} \subseteq \ldots \subseteq X^{(n)}$, constructed by pairwise disjoint open cells $c \subset X$ with the Hausdorff topology $\cup^{c \in \mathcal{C}^{\mathcal{C}}}$, such that $X^{(n)}$ is obtained from $X^{(n-1)}$ by attaching adjacent n-cells to each $(n-1)$-cell and $X^{(-1)} = \emptyset$. The respective subspaces $X^{(n)}$ are called the n-skeletons of X. A CW complex can be understood as a set of cells which are glued together at their subcells. It generalizes the concept of a graph by adding cells of dimension greater than 1.

Up to now, the definition of a cell was just based on a homeomorphism of the underlying space X and $\mathbb{R}^n$. Note that a cell does not need to be "straight," such that, e.g., a two-cell may be constructed from a single vertex and an edge connecting the vertex to itself, as, e.g., illustrated by J. Hart [34]. Alternative approaches toward the definition of cells are more restrictively based on isometry to Euclidean space, defining the notion of "convexity" first. However, it is recommendable to avoid the assumption of Euclidean space and treating the topological properties of a mesh purely based on its combinatorial relationships.

Ontological Scheme and Seven-Level Hierarchy

The concept of the fiber bundle data model builds on the paradigm that numerical data sets occurring for scientific visualization can be formulated as trivial fiber bundles (see section "Vector and Fiber Bundles"). Hence, data sets may be distinguished by their properties in the base space and the fiber space. At each point of the – discretized – base space, there are some data in the fiber space attached. Basically a fiber bundle is a set of points with neighborhood information attached to each of them. An n-dimensional array is a very simple case of a fiber bundle with neighborhood information given implicitly.

The structure of the base space is described as a CW complex, which categorizes the topological structure of an n-dimensional base space by a sequence of k-dimensional skeletons, with $0 < k < n$. These skeletons carry certain properties of the data set: the zero-skeleton describes vertices, the one-skeleton refers to edges, two-skeleton to the faces, etc., of some mesh (a triangulation of the base space). Structured grids are triangulations with implicitly given topological properties. For instance, a regular n-dimensional grid is one where each point has 2^n neighbors.

The structure of the fiber space is (usually) not discrete and given by the properties of the geometrical object residing there, such as a scalar, vector, covector, and tensor. Same as the base space, the fiber space has a specific dimensionality, though the dimensionality of the base space and fiber space is independent. Figure 3 demonstrates example images from scientific visualization classified via their fiber bundle structure. If the fiber space has vector space properties, then the fiber bundle is a vector bundle and vector operations can be performed on the fiber space, such as addition, multiplication, and derivation.

The distinction between base space and fiber space is not common use in computer graphics, where topological properties (base space) are frequently intermixed with geometrical properties (coordinate representations). Operations in the fiber space can, however, be formulated independently from the base space, which leads to a more reusable design of software components. Coordinate information, formally part of the base space, can as well be considered as fiber, leading to further generalization. The data sets describing a fiber are ideally stored as contiguous arrays in memory or disk, which allows for optimized array and vector operations. Such a storage layout turns out to be particularly useful for communicating data with the GPU using vertex buffer objects: the base space is given by vertex arrays (e.g., OpenGL *glVertexPointer*), and fibers are attribute arrays (e.g., OpenGL *glVertexAttribPointer*), in the notation of computer graphics. While the process of hardware rendering in its early times had been based on procedural descriptions (cached in display lists), vertex buffer objects are much faster in state-of-the-art technology. Efficient rendering routines are thus implemented as *maps* from fiber bundles in RAM to fiber bundles in GPU memory (eventually equipped with a GPU shader program).

A complex data structure (such as some color-coded time-dependent geometry) will be built from many data arrays. The main question that needs to be answered by

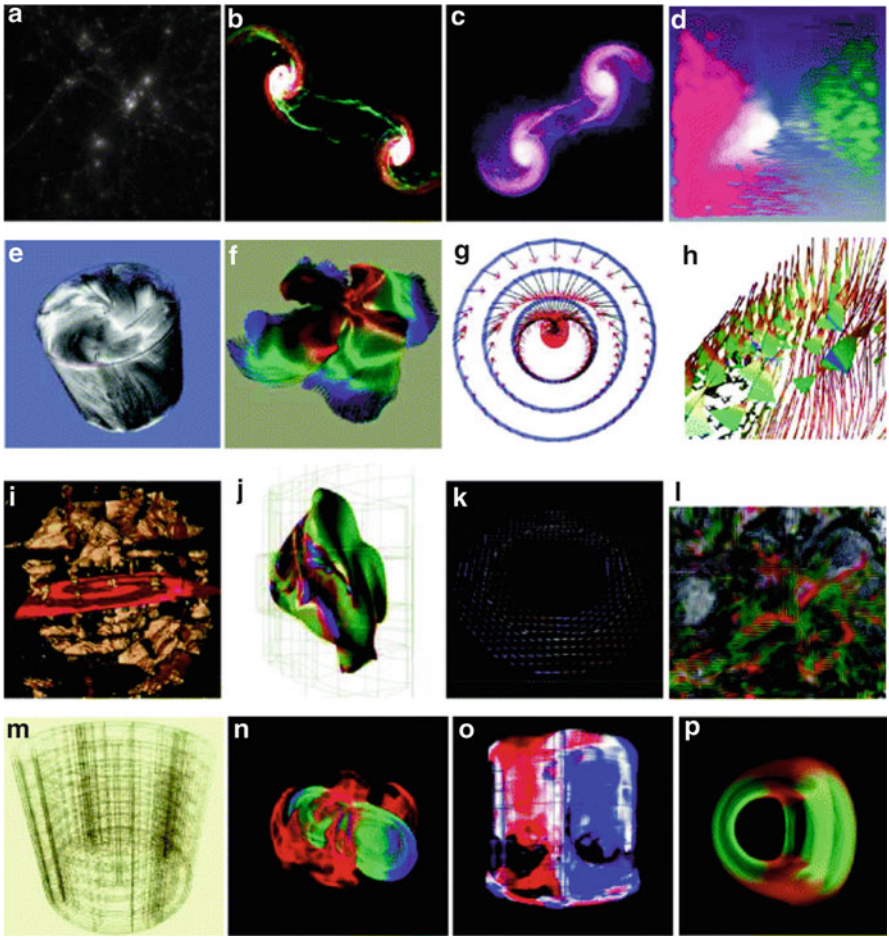


Fig. 3 Fiber bundle classification scheme for visualization methods:dimensionality of the base space (involving the k-skeleton of the discretized manifold) and dimensionality of the fiber space (involving the number of field quantities per element, zero referring to display of the mere topological structure). (**a**) Zero-cells, 0D. (**b**) Zero-cells, 1D. (**c**) Zero-cells, 3D. (**d**) Zero-cells, 6D. (**e**) One-cells, 0D. (**f**) One-cells, 1D. (**g**) One-cells, 3D. (**h**) One-cells, 6D. (**i**) Two-cells, 0D. (**j**) Two-cells, 1D. (**k**) Two-cells, 3D. (**l**) Two-cells, 6D. (**m**) Three-cells, 0D. (**n**) Three-cells, 1D. (**o**) Three-cells, 3D. (**p**) Three-cells, 6D

a data model is how to assign a semantic meaning to each of these data arrays – what do the numerical values actually *mean*? It is always possible to introduce a set of keywords with semantics attached to them. In addition, the introduction of keywords also reduces the number of possible identifiers available for user-specific purpose. This problem is also known as "name space pollution". The approach followed in the data model presented in [7] is to avoid use of keywords as much as possible. Instead, it assigns the semantics of an element of the data structure into the placement of

this element. The objective is to describe all data types that occur in an algorithm (including file reader and rendering routines) within this model. It is formulated as a graph of up to seven levels (two of them optional). Each level represents a certain property of the entire data set, the *Bundle*. These levels are called:

1. *Slice*
2. *Grid*
3. *Skeleton*
4. *Representation*
5. *Field*
6. (Fragment)
7. (Compound Elements)

Actual data arrays are stored only below the "*Field*" level. Given one hierarchy level, the next one is accessed via some identifier. The type of this identifier differs for each level: numerical values within a *Skeleton* level are grouped into *Representation* objects, which hold all information that is *relative* to a certain "representer." Such a representer may be a coordinate object that, for instance, refers to some Cartesian or polar chart, or it may well be another *Skeleton* object, either within the same *Grid* object or even within another one. An actual data set is described through the existence of entries in each level. Only two of these hierarchy levels are exposed to the end user; these are the *Grid* and *Field* levels. Their corresponding textual identifiers are arbitrary names specified by the user.

Hierarchy object	Identifier type	Identifier semantic
Bundle	Floating point number	Time value
Slice	String	Grid name
Grid	Integer set	Topological properties
Skeleton	Reference	Relationship map
Representation	String	Field name
Field	Multidimensional index	Array index

A *Grid* is subset of data within the Bundle that refers to a specific geometrical entity. A *Grid* might be a mesh carrying data such as a triangular surface, a data cube, a set of data blocks from a parallel computation, or many other data types. A *Field* is the collection of data sets given as numbers on a specific topological component of a Grid, for instance, floating point values describing pressure or temperature on a Grid's vertices. All other levels of the data model describe the properties of the Bundle as construction blocks. The usage of these construction blocks constitutes a certain language to describe data sets. A Slice is identified by a single floating point number representing time (generalization to arbitrary-dimensional parameter spaces is possible). A *Skeleton* is identified by its dimensionality, index depth (relationship to the vertices of a Grid), and refinement

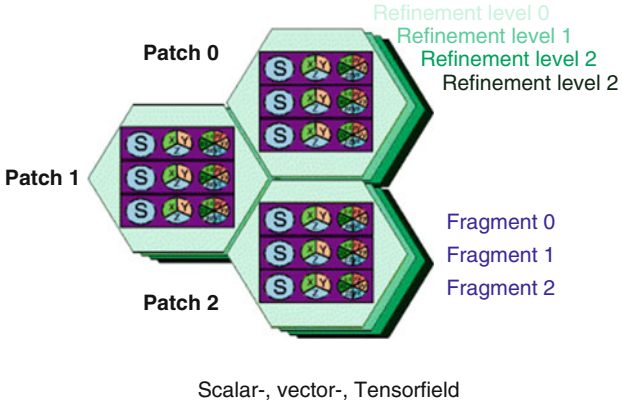


Fig. 4 Hierarchical structure of the data layout of the concept of a *field* in computer memory: (1) organization by multiple resolutions for same spatial domain; (2) multiple coordinate systems covering different spatial domains (arbitrary overlap possible); (3) fragmentation of fields into blocks (recombination from parallel data sources); and (4) layout of compound fields as components for performance reasons, indicated as S (scalar field), $\{x, y, z\}$ for vector fields, and $\{xx, xy, yy, yz, zz, zx\}$ for tensor fields

level. This will be explained in more detail in section "Topological Skeletons." The scheme also extends to cases beyond the purely mathematical basis to also cover data sets that occur in praxis, which is described in section "Non-topological representations." A representation is identified via some reference object, which may be some coordinate system or another *Skeleton*. The lowest levels of fragments and compounds describe the internal memory layout of a *Field* data set and are optional; some examples are described in [8, 9].

Field Properties

A specific *Field* identifier may occur in multiple locations. All these locations together define the properties of a field. The following four properties are expressible in the data model:

1. *Hierarchical ordering*: For a certain point in space, there exist multiple data values, one for each *refinement* level. This property describes the topological structure of the base space.
2. *Multiple coordinate systems*: One spatial point may have multiple data representations relating to different coordinate systems. This property describes the geometrical structure of the base space.
3. *Fragmentation*: Data may stem from multiple sources, such as a distributed multiprocess simulation. The field then consists of multiple data blocks, each of them covering a subdomain of the field's base space. Such field fragments may also overlap, known as "ghost zones."
4. *Separated Compounds*: A compound data type, such as a vector or tensor, may be stored in different data layouts since applications have their own preferences. An

array of tensors may also be stored as a tensor of arrays, e.g., *XYZXYZXYZXYZ* as *XXXXYYYYZZZZ*. This property describes the internal structure of the fiber space.

All of these properties are optional. In the most simple case, a field is just represented by an array of native data types; however, in the most general case (which the visualization algorithm must always support), the data are distributed over several such property elements and built from many arrays. With respect to quick transfer to the GPU, only the ability to handle multiple arrays per data set is of relevance.

Figure 4 illustrates the organization of the last four levels of the data model. These consist of Skeleton and Representation objects with optional fragmentation and compound levels. The ordering of these levels is done merely based on their semantic importance, with the uppermost level (1) embracing multiple resolutions of the spatial domain being the most visible one to the end user. Each of these resolution levels may come with different topological properties, but all arrays within the same resolution are required to be topologically compatible (i.e., share the same number of points). There might still be multiple coordinate representations required for each resolution, which constitutes the second hierarchy level (2) of multiple coordinate patches. Data per patch may well be distributed over various fragments (3), which is considered an internal structure of each patch, due to parallelization or numerical issues, but not fundamental to the physical setup. Last but not least, fields of multiple components such as vector or tensor fields may be separated into distinct arrays themselves [7]. This property, merely a performance issue of in-memory data representation, is not what the end user usually does not want to be bothered with and is thus set as the lowest level in among these four entries.

Topological Skeletons

The *Skeleton* level of the fiber bundle hierarchy describes a certain topological property. This can be the vertices, the cells, the edges, etc. Its primary purpose is to describe the skeletons of a CW complex, but they may also be used to specify mesh refinement levels and agglomerations of certain elements. All data fields that are stored within a *Skeleton* level provide the same number of elements. In other words they share their index space (a data space in HDF5 terminology). Each *Topology* object within a *Grid* object is uniquely identified via a set of integers, which are the *dimension* (e.g., the dimension of a k-cell), *index depth* (how many dereferences are required to access coordinate information in the underlying manifold), and *refinement level* (a multidimensional index, in general). Vertices – index depth 0 – of a topological space of dimension n define a Skeleton of type $(n, 0)$. Edges are one-dimensional sets of vertex indices; therefore, their index depth is 1 and their Skeleton type is (1,1). Faces are two-dimensional sets of vertex indices, hence Skeleton type (2, 1). Cells – such as a tetrahedron or hexahedra – are described by a Skeleton type (3, 1).

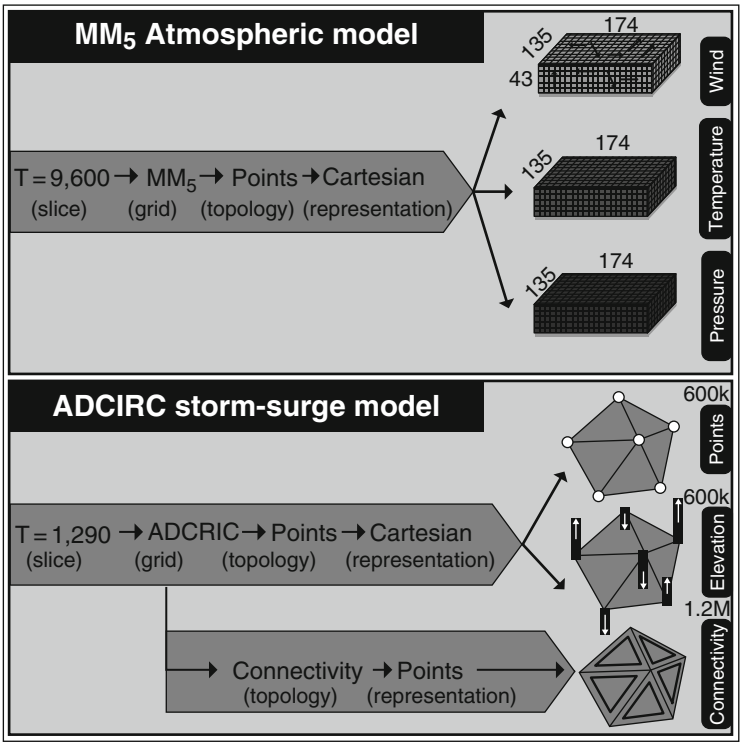


Fig. 5 The five-level organization scheme used for atmospheric data (MM5 model data) and surge data (ADCIRC simulation model), built upon common topological property descriptions with additional fields (From Venkataraman et al. [81])

All the Skeleton objects of index depth 1 build the k-skeletons of a manifold's triangulation.

Higher index depths describe sets of k-cells. For instance, a set of edges describes a line – a path along vertices in a Grid. Such a collection of edges will fit into a Skeleton of dimension 1 and depth 2, i.e., type (1, 2). It is a one-dimensional object of indices that refer to edges that refer to vertices.

Non-topological Representations

Polynomial coordinates, information on field fragments, histograms, and color maps can be formulated in the fiber bundle model as well. These quantities are no longer direct correspondences of the mathematical background, but they may still be cast into the given context.

Coordinates may be given procedurally, such as via some polynomial expression. The data for such expressions may be stored in a Skeleton of *negative* index depth – as these data are required to compute the vertex coordinates and more fundamental than these in this case.

A *fragment* of a Field given on vertices – the $(n, 0)$-Skeleton of a Grid – defines an n-dimensional subset of the Grid, defined by the hull of the vertices corresponding to the fragments. These may be expressed as a $(n, 2)$-Skeleton, where the Field object named "Positions" (represented relative to the vertices) refers to the (global) vertex indices of the respective fragments. The representation in coordinates corresponds to its range, known as the *bounding box*. Similarly, a field given on the vertices will correspond to the field's numerical *minimum/maximum range* within this fragment.

A *histogram* is the representation of a field's vertex complex in a "chart" describing the required discretization, depending on the min/max range and a number count. A *color map* (transfer function) can be interpreted as a chart object itself. It has no intrinsically geometrical meaning, but provides means to transform some data. For instance, some scalar value will be transformed to some RGB triple using some color map. A scalar field represented in a certain color map is therefore of type RGB values and could be stored as an array of RGB values for each vertex. In practice, this will not be done since such transformation is performed in real time by modern graphics hardware. However, this interpretation of a color map as a chart object tells how color maps may be stored in the fiber bundle data model.

3 Differential Forms and Topology

This section not only introduces the concepts of differential forms and their discrete counterparts but also illustrates that similar concepts are applied in several separate areas of scientific visualization. Since the available resources are discrete and finite, concepts mirroring these characteristics have to be applied to visualize complex data sets. The most distinguished algebraic structure is described by exterior algebra (or Grassmann algebra, see also section "Exterior Product"), which comes with two operations, the exterior product (or wedge product) and the exterior derivative.

Differential Forms

Manifolds can be seen as a precursor to model physical quantities of space. Charts on a manifold provide coordinates, which allows using concepts which are already well established. Furthermore, they are crucial for the field of visualization, as they are key components to obtain depictable expressions of abstract entities. Tangential vectors were already introduced in section "Tangential Vectors" as derivatives along a curve. Then a one-form α is defined as a linear mapping which assigns a value to each tangential vector v from the tangent space $T_P(M)$, i.e., $\alpha : T_P(M) \to \mathbb{R}$. They are commonly called co-variant vectors, covectors (see section "Tangential Vectors"), or Pfaff-forms. The set of one-forms generates the dual vector space or cotangential space $T_p^*(M)$. It is important to highlight that the tangent vectors $v \in T_P(M)$ are not contained in the manifold itself, so the differential forms also generate an additional space over $P \in M$. In the following, these one-forms are generalized to (alternating) differential forms.

An alternative point of view treats a tangential vector v as a linear mapping which assigns a scalar to each one-form α by $< \alpha, v > \in \mathbb{R}$. By omitting one of the arguments of the obtained mappings, $< \alpha, . >$ or $\alpha(v)$ and $<., v >$ or $v(\alpha)$, linear objects are defined. Multi-linear mappings depending on multiple vectors or covectors appear as an extension of this concept and are commonly called tensors

$$\gamma : T^{*m} \times T^n \to \mathbb{R} \tag{26}$$

where n and m are natural numbers and T^n and T^{*m} represent the n and m powered Cartesian product of the tangential space or the dual vector space (cotangential space). A tensor γ is called an (n, m)-tensor which assigns a scalar value to a set of m covectors and n vectors. All tensors of a fixed type (n, m) generate a tensor space attached at the point $P \in M$. The union of all tensor spaces at the points $P \in M$ is called a *tensor bundle*. The tangential and cotangential bundles are specialized cases for (1, 0) and (0, 1) tensor bundles, respectively. Fully antisymmetric tensors of type (0, m) may be identified with *differential forms of degree m*. For $m >$ dim(M), where dim(M) represents the dimension of the manifold, differential forms vanish.

The *exterior derivative* or Cartan derivative of differential forms generates a $p+$ 1-form df from a p-form f and conforms to the following requirements:

1. Compatibility with the wedge product (product rule): $d(\alpha \wedge \beta) = d\alpha \wedge \beta + (-1)^m \alpha \wedge d\beta$
2. Nilpotency of the operation d, $d \circ d = 0$, depicted in Fig. 11
3. Linearity

   A subset of one-forms is obtained as a differential df of zero-forms (functions) f at P and are called *exact differential forms*. For an n-dimensional manifold M, a one-form can be depicted by drawing $(n-1)$-dimensional surfaces, e.g., for the three-dimensional space, Fig. 6 depicts a possible graphical representation of a one-form attached to M. This depiction also enables a graphical representation on how to integrate differential forms, where only the number of surfaces which are intersected by the integration domain has to be counted:

$$< df, v >= df(v) = \alpha(v) \tag{27}$$

A consequence of being exact includes the closeness property $d\alpha = 0$. Furthermore, the integral $\int_{C_p} df$ with C_p representing an integration domain, e.g., an interval x_1 and x_2, results in the same value $f(x_2) - f(x_1)$. In the general case, a p-form is not always the exterior derivative of a p-one-form; therefore, the integration of p-forms is not independent of the integration domain. An example is given by the exterior derivative of a p-form β resulting in a $p+1$-form $\gamma = d\beta$. The structure of such a generated differential form can be depicted by a tube-like structure such as in Fig. 7. While the wedge product of an r-form and an s-form results in an $r + s$-form, this resulting form is not necessarily representable as a derivative. Figure 7 depicts a two-form which is not constructed by the exterior

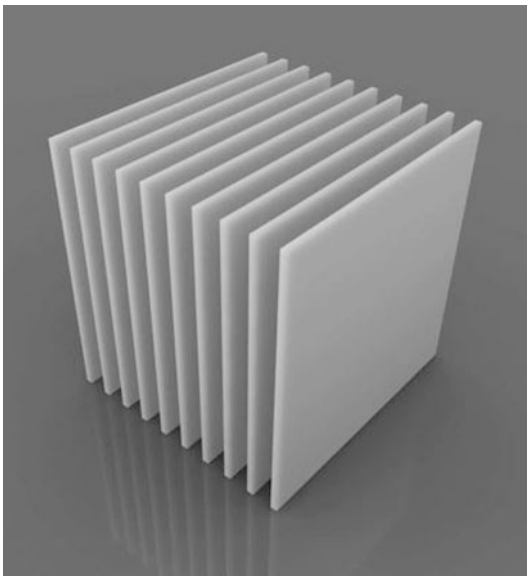


Fig. 6 Possible graphical representation of the topological structure of one-forms in three dimensions. Note that the graphical display of differential forms varies in different dimension and does not depend on the selected basis elements

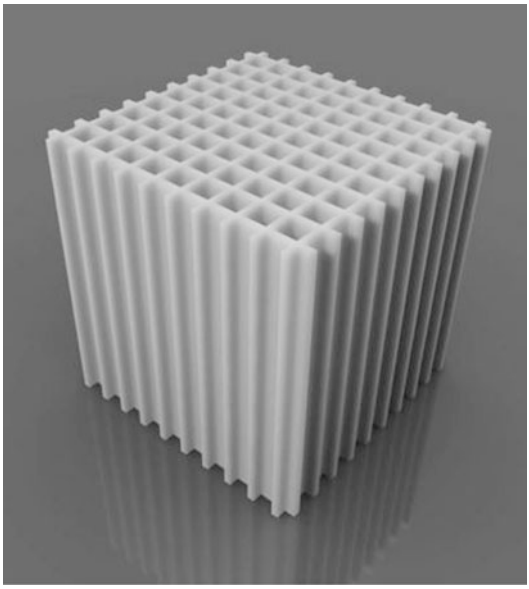


Fig. 7 Possible graphical representation of a general two-form generated by $\alpha \wedge \beta$, where α and β are one-forms. The topologically tube-like structure of the two-forms is enclosed by the depicted planes

derivative but instead by $\alpha \wedge \beta$, where α and β are one-forms. In the general case, a p-form attached on an n-dimensional manifold M is represented by using $(n - p)$-dimensional surfaces.

By sequentially applying the operation d to $(0, m)$ for $0 \leq m \leq \dim(M)$, the *de Rham complex* is obtained, which enables the investigation of the relation of closed and exact forms. The de Rham complex enables the transition from

the continuous differential forms to the discrete counterpart, so-called cochains. The already briefly mentioned topic of integration of differential forms is now mapped onto the integration of these cochains. To complete the description, the notion of chains, also modeled by multivectors (as used in Geometric Algebra, see section "Geometric Algebra" and Sect. 4) or fully antisymmetric $(n, 0)$-tensors, as description of integration domains is presented, where a chain is a collection of n-cells.

The connection between chains and cochains is investigated in algebraic topology under the name of homology theory, where chains and cochains are collected in additive Abelian groups $C_p\,(M)$.

Chains

The de Rham complex collects cochains similar to a cell complex aggregating cells as elements of chains. To use these elements, e.g., all edges, in a computational manner, a mapping of the n-cells onto an algebraic structure is needed. An algebraic representation of the assembly of cells, an n-chain, over a cell complex $\mathcal{K}$ and a vector space $\mathcal{V}$ can be written by

$$c_n = \sum\nolimits_{i=1}^{j} w_i \tau_n^i \quad \tau_n^i \in \mathcal{K},\ w_i \in \mathcal{V}$$

which is closed under reversal of the orientation:

$$\forall \tau_n^i \in c_n \quad \text{there is} - \tau_n^i \in c_n$$

The different topological elements are called cells, and the dimensionality is expressed by adding the dimension such as a three-cell for a volume, a two-cell for surface elements, a one-cell for lines, and a zero-cell for vertices. If the coefficients are restricted to $\{-1, 0, 1\} \in \mathbb{Z}$, the following classification for elements of a cell complex is obtained:

- 0: if the cell is not in the complex
- 1: if the unchanged cell is in the complex
- -1 : if the orientation is changed

The so-called boundary operator is a map between sets of chains C_p on a cell complex K. Let us denote the ith p-cell as $\tau_p^i = k_0, \ldots k_p$, whereby $\tau_p^i \in K$. The boundary operator ∂_p defines a $(p-1)$-chain computed from a p-chain: $\partial_p : C_p(K) \dashrightarrow C_{p-1}(K)$. The boundary of a cell τ_p^j can be written as alternating sum over elements of dimension $p-1$:

$$\partial_p \tau_p^i = \sum\nolimits_i (-1)^i [k_0, k_1, \ldots, \tilde{k_i}, \ldots k_n] \tag{28}$$

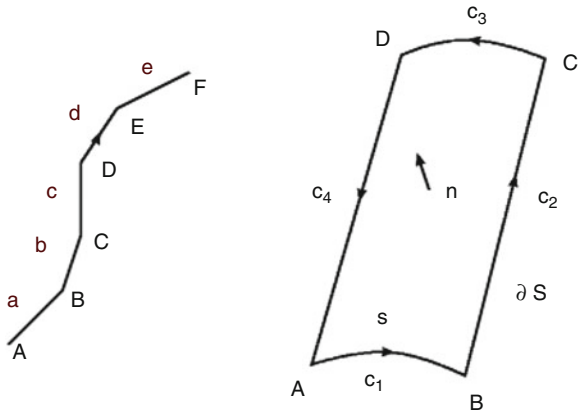


Fig. 8 Representation of a one-chain τ_1^i with zero-chain boundary τ_0^j (*left*) and a two-chain τ_2 with one-chain boundary τ_1^k (*right*)

where $\tilde{k_i}$ indicates that k_i is deleted from the sequence. This map is compatible with the additive and the external multiplicative structure of chains and builds a linear transformation:

$$C_p \rightarrow C_{p-1} \tag{29}$$

Therefore, the boundary operator is linear

$$\partial \left(\sum_i w_i \tau_p^i \right) = \sum_i w_i \left(\partial \tau_p^i \right) \tag{30}$$

which means that the boundary operator can be applied separately to each cell of a chain. Using the boundary operator on a sequence of chains of different dimensions results in a chain complex $C_* = \{C_p, \partial_p\}$ such that the complex property

$$\partial_{p-1} \partial_p = 0 \tag{31}$$

is given. Homological concepts are visible here for the first time, as homology examines the connectivity between two immediately neighboring dimensions. Figure 8 depicts two examples of one-chains and two-chains and an example of the boundary operator.

Applying the appropriate boundary operator to the two-chain example reads

$$\begin{gathered} \partial_2 \tau_2 = \tau_1^1 + \tau_1^2 + \tau_1^3 + \tau_1^4 \\ \partial_1(\tau_1^1 + \tau_1^2 + \tau_1^3 + \tau_1^4) = \tau_0^1 + \tau_0^2 - \tau_0^2 + \tau_0^3 - \tau_0^3 + \tau_0^4 - \tau_0^4 - \tau_0^1 = 0 \end{gathered} \tag{33}$$

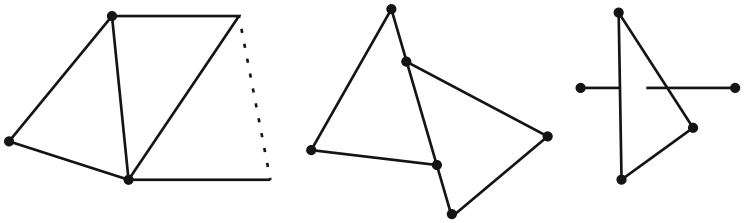


Fig. 9 Examples of violations of correct cell attachment. *Left*: missing zero-cell. *Middle*: cells do not intersect at vertices. *Right*: intersection of cells

A different view on chain complexes presents itself when the main focus is placed on the cells within a chain. To cover even the most abstract cases, a cell is defined as a subset $c \subset X$ of a Hausdorff space X if it is homeomorphic to the interior of the open n-dimensional ball $\mathbb{D}^n = \{x \in \mathbb{R}^n : |x| < 1\}$. The number n is unique due to the *invariance of domain* theorem [13] and is called the dimension of c, whereas homeomorphic means that two or more spaces share the same topological characteristics. The following list assigns terms corresponding to other areas of scientific computing:

- 0-cell: point
- 1-cell: edge
- 2-cell: facet
- n-cell: cell

A cell complex $\mathcal{K}$ (see also section "Topology: Discretized Manifolds") can be described by a set of cells that satisfy the following properties:

- The boundary of each p-cell τ_p^i is a finite union of $(p-1)$-cells in $\mathcal{K} : \partial_p \tau_p^i = \cup^m \tau_{p-1}^m$.
- The intersection of any two cells τ_p^i, τ_p^j in $\mathcal{K}$ is either empty or is a unique cell in $\mathcal{K}$.

The result of these operations are subspaces $X^{(n)}$ which are called the n-skeletons of the cell complex. Incidence and adjacence relations are then available. Examples for incidence can be given by vertex on edge relation and for adjacency by vertex to vertex relations. This cell complex with the underlying topological space guarantees that all interdimensional objects are connected in an appropriate manner. Although there are various possible attachments of cells, only one process results in a cell complex, see Fig. 9.

Cochains

In addition to chain and cell complices, scientific visualization requires the notation and access mechanisms to global quantities related to macroscopic n-dimensional

space-time domains. The differential forms which are necessary concepts to handle physical properties can also be projected onto discrete counterparts, which are called *cochains*. This collection of possible quantities, which can be measured, can then be called a section of a fiber bundle, which permits the modeling of these measurements as a function that can be integrated on arbitrary n-dimensional (sub)domains or multivectors. This function can then be seen as the abstracted process of measurement of this quantity [55, 75]. The concept of cochains allows the association of numbers not only to single cells, as chains do, but also to assemblies of cells. Briefly, the necessary requirements are that this mapping is not only orientation dependent but also linear with respect to the assembly of cells. A cochain representation is now the global quantity association with subdomains of a cell complex, which can be arbitrarily built to discretize a domain.

A linear transformation σ of the n-chains into the field $\mathbb{R}$ of real numbers forms a vector space $c_n \longrightarrow \wedge\{\sigma\}\mathbb{R}$ and is called a vector-valued m-dimensional cochain or short m-cochain. The coboundary δ of an m-cochain is an $(m+1)$-cochain defined as

$$\delta c^m = \sum\nolimits_i v_i \tau_i, \quad \text{where} \quad v_i = \sum\nolimits_{b \in \text{ faces}(\tau_i)} \sigma(b, \tau_i) c_m(b) \tag{34}$$

Thus, the coboundary operator assigns nonzero coefficients only to those $(m+1)$ cells that have c_m as a face. As can be seen, δc_m depends not only on c_m but on how c_m lies in the complex $\mathcal{K}$. This is a fundamental difference between the two operators ∂ and δ. An example is given in Fig. 10 where the coboundary operator is used on a one-cell. The right part $\delta \circ \delta\mathcal{K}$ of Fig. 10 is also depicted for the continuous differential forms in Fig. 7. The coboundary of an m-cochain is an $(m+1)$-cochain which assigns to each $(m+1)$ cell the sum of the values that the $m+1$-cochain assign to the m-cells which form the boundary of the $(m+1)$ cell. Each quantity appears in the sum multiplied by the corresponding incidence number. Cochain complices [33, 35] are similar to chain complices except that the arrows are reversed, so a cochain complex $C^* = \{C^m, \delta^m\}$ is a sequence of modules C^m and homomorphisms:

$$\delta^m : C^m \to C^{m+1} \tag{35}$$

such that

$$\delta^{m+1}\delta^m = 0 \tag{36}$$

$\mathcal{K}^1 - \overset{\delta}{\to} \delta\mathcal{K}^1 - \overset{\delta}{\to} \delta \circ \delta\mathcal{K}^1 = 0$. Proceeding from left to right, a one-cochain represented by a line segment, a two-cochain generated by the product of two one-forms, and a three-cochain depicted by volume objects are illustrated.

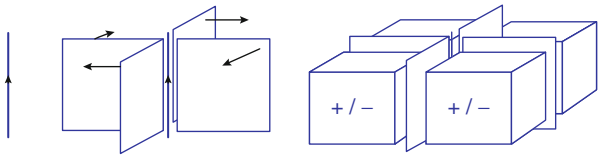


Fig. 10 Cochain complex with the corresponding coboundary operator

Then, the following sequence with $\delta \circ \delta = 0$ is generated:

$$0 - \stackrel{\delta}{\rightarrow} C^0 - \stackrel{\delta}{\rightarrow} C^1 - \stackrel{\delta}{\rightarrow} C^2 - \stackrel{\delta}{\rightarrow} C^3 - \stackrel{\delta}{\rightarrow} 0 \tag{37}$$

Cochains are the algebraic equivalent of alternating differential forms, while the coboundary process is the algebraic equivalent of the external derivative and can therefore be considered as the discrete counterpart of the differential operators:

- grad.
- curl.
- div.

It indeed satisfies the property $\delta \circ \delta \equiv 0$ corresponding to

- curlgrad. $\equiv 0$
- divcurl. $\equiv 0$

Duality Between Chains and Cochains

Furthermore, a definition of the adjoint nature of $\partial, \delta : C^p \rightarrow C^{p+1}$ can be given:

$$\langle c^p, \partial c_{p+1} \rangle = \langle \delta c^p, c_{p+1} \rangle \tag{38}$$

The concepts of chains and cochains coincide on finite complices [45]. Geometrically, however, C_p and C^p are distinct [12] despite an isomorphism. An element of C_p is a formal sum of p-cells, where an element of C^p is a linear function that maps elements of C_p into a field. Chains are dimensionless multiplicities of aggregated cells, whereas those associated with cochains may be interpreted as physical quantities [65]. The extension of cochains from single cell weights to quantities associated with assemblies of cells is not trivial and makes cochains very different from chains, even on finite cell complices. Nevertheless, there is an important duality between p-chains and p-cochains. The first part of the de Rham (cohomology group) complex, depicted in Fig. 11 on the left, is the set of closed one-forms modulo the set of exact one-forms denoted by

$$H^1 = Z^1 / B^1 \tag{39}$$

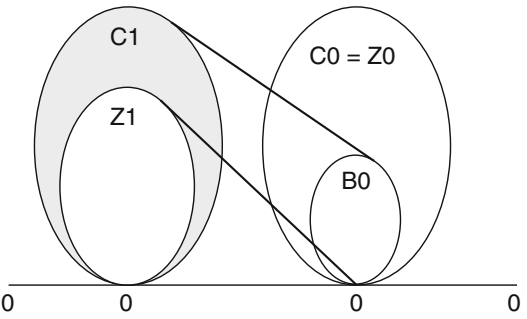


Fig. 11 A graphical representation of closed and exact forms. The forms Z_1, B_1, Z_0, and B_0 are closed forms, while only the forms B_0 and B_1 are exact forms. The nilpotency of the operation d forces the exact forms to vanish

This group is therefore trivial (only the zero element) if all closed one-forms are exact. If the corresponding space is multiply connected, then there are closed one-chains that are not themselves boundaries, and there are closed one-forms that are not themselves exact.

For a chain $c_p \in C_p(\mathcal{K}, \mathbb{R})$ and a cochain $c^p \in C^p(\mathcal{K}, \mathbb{R})$, the integral of c^p over c_p is denoted by $\int_{c_p} c^p$, and integration can be regarded as a mapping, where D represents the corresponding dimension:

$$\int : C_p(\mathcal{K}) \times C^p(\mathcal{K}) \to \mathbb{R}, \qquad \text{for } 0 \leq p \leq D \tag{40}$$

Integration in the context of cochains is a linear operation: given $a_1, a_2 \in \mathbb{R}$, $c^{p,1} c^{p,2} \in C^p(\mathcal{K})$ and $c_p \in C_p(\mathcal{K})$, reads

$$\int_{c_p} a_1 c^{p,1} + a_2 c^{p,2} = a_1 \int_{c_p} c^{p,1} + a_2 \int_{c_p} c^{p,2} \tag{41}$$

Reversing the orientation of a chain means that integrals over that chain acquire the opposite sign

$$\int_{-c_p} c^p = -\int_{c_p} c^p \tag{42}$$

using the set of p-chains with vector space properties $C_p(\mathcal{K}, \mathbb{R})$, e.g., linear combinations of p-chains with coefficients in the field $\mathbb{R}$. For coefficients in $\mathbb{R}$, the operation of integration can be regarded as a bilinear pairing between p-chains and p-cochains. Furthermore, for reasonable p-chains and p-cochains, this bilinear pairing for integration is nondegenerate,

$$\text{if} \quad \int_{c_p} c^p = 0 \quad \forall c_p \in C_p(\mathcal{K}), \quad \text{then } c^p = 0 \tag{43}$$

and

$$\text{if} \quad \int_{c_p} c^p = 0 \quad \forall c^p \in C^p(\mathcal{K}), \quad \text{then } c_p = 0 \tag{44}$$

The integration domain can be described by, using Geometric Algebra notation, the exterior product applied to multivectors. An example is then given by the generalized Stokes theorem:

$$\int_{c_p} df = \int_{\partial c_p} f \tag{45}$$

or

$$< df, c_p > = < f, \partial c_p > \tag{46}$$

The generalized Stokes theorem combines two important concepts, the integration domain and the form to be integrated.

Homology and Cohomology

The concepts of chains can also be used to characterize properties of spaces, the homology and cohomology, where it is only necessary to use $C_p(\mathcal{K}, \mathbb{Z})$. The algebraic structure of chains is an important concept, e.g., to detect a p-dimensional hole that is not the boundary of a $p+1$-chain, which is called a p-cycle. For short, a cycle is a chain whose boundary is $\partial_p c_p = 0$, a closed chain. The introduced boundary operator can also be related to homological terms. A boundary is a chain b_p for which there is a chain c_p such that $\partial_p c_p = b_p$. Since $\partial \circ \partial = 0$, $B_n \subset Z_n$ is obtained. The homology is then defined by $H_n = Z_n / B_n$. The homology of a space is a sequence of vector spaces. The topological classification of homology is defined by

$$\begin{aligned} B_p &= \text{im } \partial_{p+1} \qquad \text{and} \\ Z_p &= \text{ker } \partial_p \end{aligned}$$

so that $B_p \subset Z_p$ and

$$H_p = Z_p / B_p$$

where $\beta_p = \{\text{Rank}\}\ H_p$. Here $\{\text{im}\}$ is the image and $\{\text{ker}\}$ is the kernel of the mapping. For cohomology

$$\begin{aligned} B^p &= \text{im } d^{p+1} \qquad \text{and} \\ Z^p &= \text{ker } d^p \end{aligned}$$

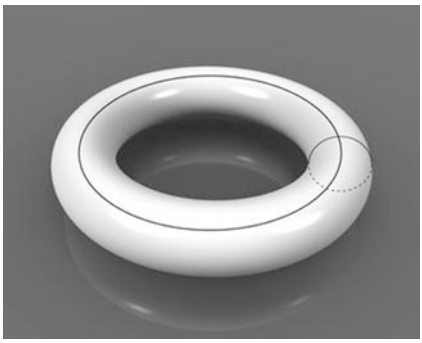


Fig. 12 Topologically a torus is the product of two circles. The partially shaded circle is spun around the fully drawn circle which can be interpreted as the closure of a cylinder onto itself

so that $B^p \subset Z^p$ and

$$H^p = Z^p / B^p$$

where $\beta^p = \{\text{Rank}\}\ H^p$. An important property of these vector spaces is given by β, which corresponds to the dimension of the vector spaces H and is called the Betti number [35, 86]. Betti numbers identify the number of nonhomologous cycles which are not boundaries:

- β_0 counts the number of connected components.
- β_1 counts the number of tunnels (topological holes).
- β_2 counts the number of enclosed cavities.

The number of connected components gives the number of distinct entities of a given object, whereas tunnels describe the number of separated parts of space. In contrast to a tunnel, the enclosed cavities are completely bounded by the object.

Examples for the Betti numbers of various geometrical objects are stated by:

- Cylinder: $\beta_0 = 1, \beta_1 = 1, \beta_n = 0 \forall n \geq 2$. The cylinder consists of one connected component, which forms a single separation of space. Therefore no enclosed cavitiy is present.
- Sphere: $\beta_0 = 1, \beta_1 = 0, \beta_2 = 1, \beta_n = 0 \forall n \geq 3$. If β_1 and β_2 are switched, a sphere is obtained by contracting the separation by generating an enclosed cavity from the tunnel.
- Torus: $\beta_0 = 1, \beta_1 = 2, \beta_2 = 1, \beta_n = 0 \forall n \geq 3$. Closing a cylinder onto itself results in a torus which not only generates an enclosed cavity but also maintains the cylinder's tunnel. An additional tunnel is introduced due to the closing procedure which is depicted in Fig. 12 as the central hole.

The Euler characteristics, which is an invariant, can be derived from the Betti numbers by: $\xi = \beta_0 - \beta_1 + \beta_2$.

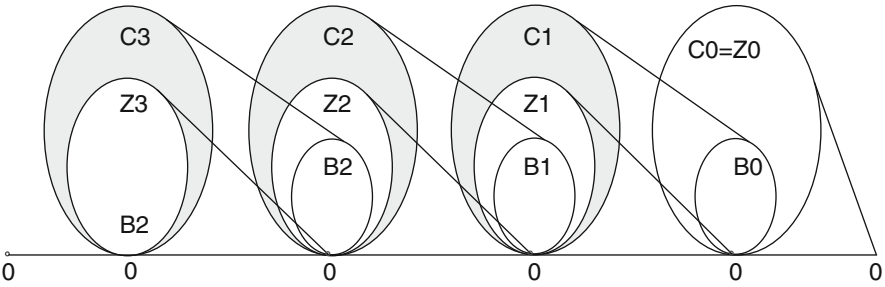


Fig. 13 A graphical representation of (co)homology for a three-dimensional cell complex

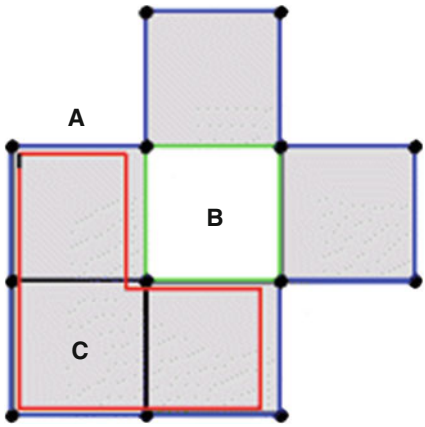


Fig. 14 Illustration of cycles A, B, C and a boundary C. A, B are not boundaries

Figure 13 depicts the homology of a three-dimensional chain complex with the respective images and kernels, where the chain complex of $\mathcal{K}$ is defined by $\{\text{im}\}\ \partial_{p+1} \subseteq \{\ker\}\ \partial_p$. As can be seen, the boundary operator expression yields $\partial_p \circ \partial_{p+1} = 0$.

To give an example, the first homology group is the set of closed one-chains (curves) modulo the closed one-chains which are also boundaries. This group is denoted by $H_1 = Z_1/B_1$, where Z_1 are cycles or closed one-chains and B_1 are one-boundaries. Another example is given in Fig. 14, where A, B, C are cycles and a boundary C, but A, B are not boundaries.

Topology

Conceptual consistency in scientific visualization is provided by topology. Cell complices convey topology in a computationally treatable manner and can therefore be introduced by much simpler definitions. A topological space $(X, \mathcal{T})$ is the collection of sets $\mathcal{T}$ that include:

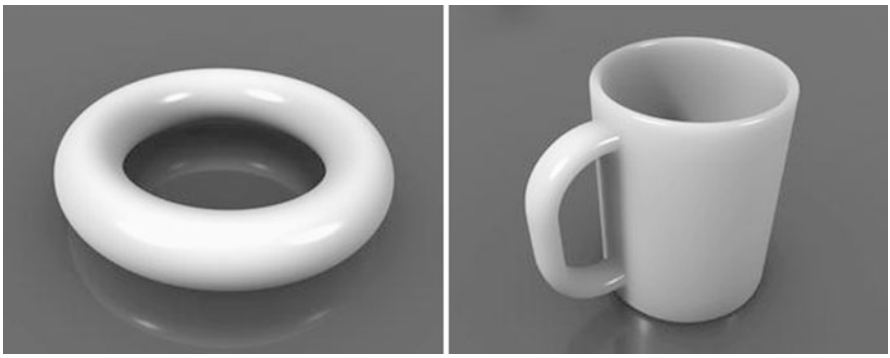


Fig. 15 Topologically a torus and a coffee mug are equivalent and so have the same Betti numbers

- The space itself X and the empty set $\emptyset$
- The union of any of these sets
- The finite intersection of any of the sets

The family $\mathcal{T}$ is called a *topology* on X, and the members of $\mathcal{T}$ are called *open sets*. As an example a basic set $X = \{a, b, c\}$ and a topology is given:

$$\begin{aligned}(X, \mathcal{T}) = \{ \quad & \emptyset, \\ & \{a\}, \{b\}, \{c\}, \\ & \{a, b\}, \{a, c\}, \{b, c\}, \\ & \{a, b, c\}\}\end{aligned}$$

The general definition for a topological space is very abstract and allows several topological spaces which are not useful in scientific visualization, e.g., a topological space $(X, \mathcal{T})$ with a trivial topology $\mathcal{T} = \{\emptyset, X\}$. So basic mechanisms of separation within a topological space are required, e.g., the Hausdorff property. A topological space $(X, \mathcal{T})$ is said to be Hausdorff if, given $x, y \in X$ with $x \neq y$, there exist open sets U_1, U_2 such that $x \in U_1, y \in U_2$ and $U_1 \cap U_2 = \emptyset$. But the question remains on what "topology" actually is. A brief explanation is given by the study of properties of an object that do not change under *deformation*. To describe this deformation process, abstract rules can be stated and if they are true, then an object A can be transformed into an object B without change. The two objects A, B are then called homeomorphic:

- All points of $A \leftrightarrow$ all points of B
- $1 - 1$ correspondence (no overlap)
- Bicontinous (continuous both ways)
- Cannot tear, join, poke/seal holes

The deformation is $1 - 1$ if each point of A maps to a single point on B and there is no overlap. If this deformation is continuous, A cannot be teared, joined, disrupted, or sealed up. If two objects are homeomorphic, then they are topologically equivalent. Figure 15 illustrates an example of a torus and coffee mug which are a prominent example for topological equivalence. The torus can be continuously deformed, without tearing, joining, disrupting, or sealing up, into a cup. The hole in the torus becomes the handle of the cup. But why should anybody in visualization be concerned about how objects can be deformed? Topology is much more than the illustrated properties, it can be much better described by the study of connectedness:

- Understanding of space properties: how connectivity happens.
- Analysis of space properties: how connectivity can be determined.
- Articulation of space properties: how connectivity can be described.
- Control about space properties: how connectivity can be enforced.

Topology studies properties of sets that do not change under well-behaved transformations (homeomorphisms). These properties include completeness and compactness. In visualization, one property is of significance: connectedness. Especially, how many disjoint components can be distinguished and how many holes (or tunnels) are in these components. Geometric configuration is another interesting aspect in visualization because it is important to know which of these components have how many holes, and where the holes are relative to each other. Several operations in scientific visualization can be summarized:

- Simplification: reduction of data complexity. If objects are described with fewer properties, important properties such as components or holes should be retained or removed, if these properties become insignificant, unnecessary, or imperceptible.
- Compression: reduction of data storage. It is important that each operation does not alter important features (interaction of geometrical and topological features).
- Texturing: visualization context elements. How can a texture kept consistent if an object, e.g., a torus, is transformed into another object, e.g., a coffee cup.
- Morphing: transforming one object into another. If an object is morphed into another, topological features have to remain, e.g., the torus hole has to become the coffee cup handle hole.

4 Geometric Algebra Computing

Geometric Algebra as a general mathematical system unites many mathematical concepts such as vector algebra, quaternions, Plücker coordinates, and projective geometry, and it easily deals with geometric objects, operations, and transformations. A lot of applications in computer graphics, computer vision, and other engineering areas can benefit from these properties. In a ray-tracing application, for

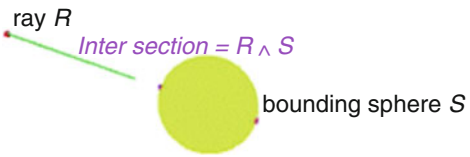


Fig. 16 Spheres and lines are basic entities of Geometric Algebra to compute with. Operations like the intersection of them are easily expressed with the help of their outer product. The result of the intersection of a ray and a (bounding) sphere is another geometric entity, the point pair of the two points of the line intersecting the sphere. The sign of the square of the point pair easily indicates whether there is a real intersection or not

instance, the intersection of a ray and a bounding sphere is needed. According to Fig. 16, this can be easily expressed with the help of the outer product of these two geometric entities.

Geometric Algebra is based on the work of Hermann Grassmann (see the conference [62] celebrating his 200th birthday in 2009) and William Clifford [20, 21]. Pioneering work has been done by David Hestenes, who first applied Geometric Algebra to problems in mechanics and physics [39, 40].

The first time Geometric Algebra was introduced to a wider computer graphics audience was through a couple of courses at the SIGGRAPH conferences in 2000 and 2001 (see [57]) and later at the Eurographics [41]. Researchers at the University of Cambridge, UK, have applied Geometric Algebra to a number of graphics-related projects. Geomerics [71] is a start-up company in Cambridge specializing in simulation software for physics and lighting, which presented its new technology allowing real-time radiosity in videogames utilizing commodity graphics-processing hardware. The technology is based on Geometric Algebra wavelet technology. Researchers at the University of Amsterdam, the Netherlands, are applying their fundamental research on Geometric Algebra to 3D computer vision and to ray tracing and on the efficient software implementation of Geometric Algebra. Researchers from Guadalajara, Mexico, are primarily dealing with the application of Geometric Algebra in the field of computer vision, robot vision, and kinematics. They are using Geometric Algebra, for instance, for tasks like visual-guided grasping, camera self-localization, and reconstruction of shape and motion. Their methods for geometric neural computing are used for tasks like pattern recognition [5]. Registration, the task of finding correspondences between two point sets, is solved based on Geometric Algebra methods in [65]. Some of their kinematics algorithms are dealing with inverse kinematics, fixation, and grasping as well as with kinematics and differential kinematics of binocular robot heads. At the University of Kiel, Germany, researchers are applying Geometric Algebra to robot vision and pose estimation [66]. They also do some interesting research dealing, for instance, with neural networks based on Geometric Algebra [14]. In addition to these examples, there are many other applications like Geometric Algebra Fourier transforms for the visualization and analysis of vector fields [24] or classification and clustering of spatial patterns with Geometric Algebra [63] showing the wide area

Table 1 Multiplication table of the 2D Geometric Algebra. This algebra consists of basic algebraic objects of grade (dimension) 0, the scalar; of grade 1, the two basis vectors e_1 and e_2; and of grade 2, the bi-vector $e_1 \wedge e_2$, which can be identified with the imaginary number i squaring to -1

	1	e_1	e_2	$e_1 \wedge e_2$
1	1	e_1	e_2	$e_1 \wedge e_2$
e_1	e_1	1	$e_1 \wedge e_2$	e_2
e_2	e_2	$-e_1 \wedge e_2$	1	$-e_1$
$e_1 \wedge e_2$	$e_1 \wedge e_2$	$-e_2$	e_1	-1

Table 2 List of the basic geometric primitives provided by the 5D conformal Geometric Algebra. The bold characters represent 3D entities ($\mathbf{x}$ is a 3D point, $\mathbf{n}$ is a 3D normal vector, and $\mathbf{x}^2$ is the scalar product of the 3D vector $\mathbf{x}$). The two additional basis vectors e_0 and e_∞ represent the origin and infinity. Based on the outer product, *circles* and *lines* can be described as intersections of two spheres, respectively two planes. The parameter r represents the radius of the sphere and the parameter d the distance of the plane to the origin

Entity	Representation
Point	$P = \mathrm{x} + \frac{1}{2}\mathrm{x}^2 e_\infty + e_0$
Sphere	$S = P - \frac{1}{2} r^2 e_\infty$
Plane	$\pi = \mathbf{n} + d e_\infty$
Circle	$Z = S_1 \wedge S_2$
Line	$L = \pi_1 \wedge \pi_2$

of possibilities of advantageously using this mathematical system in engineering applications.

Benefits of Geometric Algebra

As follows, we highlight some of the properties of Geometric Algebra that make it advantageous for a lot of engineering applications.

Unification of Mathematical Systems

In the wide range of engineering applications, many different mathematical systems are currently used. One notable advantage of Geometric Algebra is that it subsumes mathematical systems like vector algebra, complex analysis, quaternions, or Plücker coordinates. Table 1, for instance, describes how complex numbers can be identified within the 2D Geometric Algebra. This algebra does not only contain the two basis vectors e_1 and e_2 but also basis elements of grade (dimension) 0 and 2 representing the scalar and imaginary part of complex numbers.

Other examples are Plücker coordinates based on the description of lines in conformal geometric algebra (see section "Conformal Geometric Algebra") or quaternions as to be identified in Fig. 19 with their imaginary units.

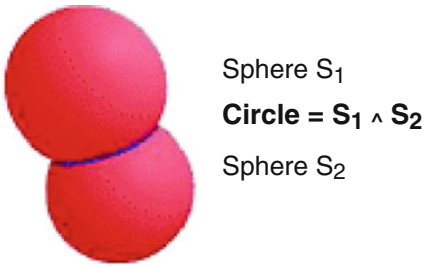


Fig. 17 Spheres and circles are basic entities of Geometric Algebra. Operations like the intersection of two spheres are easily expressed

Uniform Handling of Different Geometric Primitives

Conformal Geometric Algebra, the Geometric Algebra of conformal space we focus on, is able to easily treat different geometric objects. Table 2 presents the representation of points, lines, circles, spheres, and planes as the same entities algebraically. Consider the spheres of Fig. 17, for instance. These spheres are simply represented by

$$S = P - \frac{1}{2} r^2 e_\infty \tag{47}$$

based on their center point P, their radius r, and the basis vector e_∞ which represents the point at infinity. The circle of intersection of the spheres is then easily computed using the outer product to operate on the spheres as simply as if they were vectors:

$$Z = S_1 \wedge S_2 \tag{48}$$

This way of computing with Geometric Algebra clearly benefits computer graphics applications.

Simplified Geometric Operations

Geometric operations like rotations, translations (see [41]), and reflections can be easily treated within the algebra. There is no need to change the way of describing them with other approaches (vector algebra, for instance, additionally needs matrices in order to describe transformations).

Figure 18 visualizes the reflection of the ray R from one plane

$$\pi = \mathbf{n} + d e_\infty \tag{49}$$

(see Table 2). The reflected line, drawn in magenta,

$$\mathbf{R}_{\text{reflected}} = -\frac{\pi \mathbf{R}}{\pi} \tag{50}$$

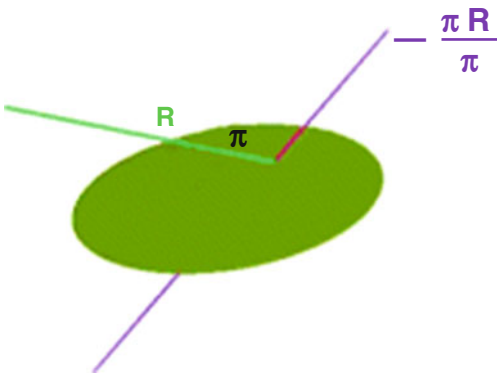


Fig. 18 The ray R is reflected from the plane π computing $-\frac{\pi \mathbf{R}}{\pi}$

is computed with the help of the reflection operation including the reflection object as well as the object to be reflected.

More Efficient Implementations

Geometric Algebra as a mathematical language suggests a clearer structure and greater elegance in understanding methods and formulae. But, what about the runtime performance for derived algorithms? Geometric Algebra inherently has a large potential for creating optimizations leading to more highly efficient implementations especially for parallel platforms. Gaalop [44], as presented in section "Computational Efficiency of Geometric Algebra Using Gaalop," is an approach offering dramatically improved optimizations.

Conformal Geometric Algebra

Conformal Geometric Algebra is a 5D Geometric Algebra based on the 3D basis vectors e_1, e_2, and e_3 as well as on the two additional base vectors e_0 representing the origin and e_∞ representing infinity.

Blades are the basic computational elements and the basic geometric entities of geometric algebras. The 5D conformal Geometric Algebra consists of blades with *grades* (dimension) 0, 1, 2, 3, 4, and 5, whereby a scalar is a *0-blade* (blade of grade 0). The element of grade five is called the pseudoscalar. A linear combination of blades is called a *k-vector*. So a bi-vector is a linear combination of blades with grade 2. Other k-vectors are vectors (grade 1), tri-vectors (grade 3), and quadvectors (grade 4). Furthermore, a linear combination of blades of different grades is called a *multivector*. Multivectors are the general elements of a Geometric Algebra. Table 3 lists all the 32 blades of conformal Geometric Algebra. The indices indicate 1, scalar; 2–6, vector; 7–16, bi-vector; 17–26, tri-vector; 27–31, quadvector; and 32, pseudoscalar.

A point $P = x_1 e_1 + x_2 e_2 + x_3 e_3 + \frac{1}{2}\mathbf{x}^2 e_\infty + e_0$ (see Table 2), for instance, can be written in terms of a multivector as the following linear combination of blades $b[i]$:

Table 3 The 32 blades of the 5D conformal Geometric Algebra

Index	Blade	Grade
1	1	0
2	e_1	1
3	e_2	1
4	e_3	1
5	e_∞	1
6	e_0	1
7	$e_1 \wedge e_2$	2
8	$e_1 \wedge e_3$	2
9	$e_1 \wedge e_\infty$	2
10	$e_1 \wedge e_0$	2
11	$e_2 \wedge e_3$	2
12	$e_2 \wedge e_\infty$	2
13	$e_2 \wedge e_0$	2
14	$e_3 \wedge e_\infty$	2
15	$e_3 \wedge e_0$	2
16	$e_\infty \wedge e_0$	2
17	$e_1 \wedge e_2 \wedge e_3$	3
18	$e_1 \wedge e_2 \wedge e_\infty$	3
19	$e_1 \wedge e_2 \wedge e_0$	3
20	$e_1 \wedge e_3 \wedge e_\infty$	3
21	$e_1 \wedge e_3 \wedge e_0$	3
22	$e_1 \wedge e_\infty \wedge e_0$	3
23	$e_2 \wedge e_3 \wedge e_\infty$	3
24	$e_2 \wedge e_3 \wedge e_0$	3
25	$e_2 \wedge e_\infty \wedge e_0$	3
26	$e_3 \wedge e_\infty \wedge e_0$	3
27	$e_1 \wedge e_2 \wedge e_3 \wedge e_\infty$	4
28	$e_1 \wedge e_2 \wedge e_3 \wedge e_0$	4
29	$e_1 \wedge e_2 \wedge e_\infty \wedge e_0$	4
30	$e_1 \wedge e_3 \wedge e_\infty \wedge e_0$	4
31	$e_2 \wedge e_3 \wedge e_\infty \wedge e_0$	4
32	$e_1 \wedge e_2 \wedge e_3 \wedge e_\infty \wedge e_0$	5

$$P = x_{1*}b[2] + x_{2*}b[3] + x_{3*}b[4] + \frac{1}{2}\mathbf{x}_*^2 b[5] + b[6] \tag{51}$$

with multivector indices according to Table 3.

Figure 19 describes some interpretations of the 32 basis blades of conformal Geometric Algebra. Scalars like the number π are grade 0 entities. They can be combined with the blade representing the imaginary unit i to complex numbers or with the blades representing the imaginary units i, j, k to quaternions. Since quaternions describe rotations, this kind of transformation can be handled within the

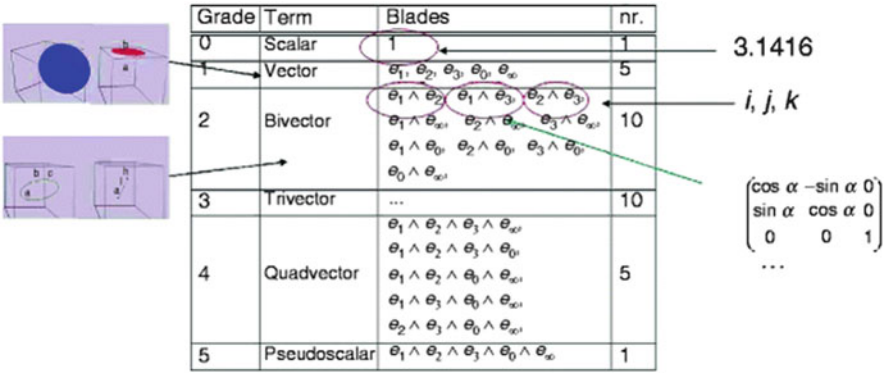


Fig. 19 The blades of conformal Geometric Algebra. *Spheres* and *planes*, for instance, are vectors. *Lines* and *circles* can be represented as bi-vectors. Other mathematical systems like complex numbers or quaternions can be identified based on their imaginary units i, j, k. This is why also transformations like rotations can be handled within the algebra

Table 4 The extended list of the two representations of the conformal geometric entities. The IPNS representations as described in Table 1 have also an OPNS representation, which are dual to each other (indicated by the star symbol). In the OPNS representation, the geometric objects are described with the help of the outer product of conformal points that are part of the objects, for instance, lines as the outer product of two points and the point at infinity

Entity	IPNS representation	OPNS representation
Point	$P = \mathbf{x} + \frac{1}{2}\mathbf{x}^2 e_\infty + e_0$	
Sphere	$S = P - \frac{1}{2}r^2 e_\infty$	$S^* = P_1 \wedge P_2 \wedge P_3 \wedge P_4$
Plane	$\pi = \mathbf{n} + d e_\infty$	$\pi^* = P_1 \wedge P_2 \wedge P_3 \wedge e_\infty$
Circle	$Z = S_1 \wedge S_2$	$Z^* = P_1 \wedge P_2 \wedge P_3$
Line	$L = \pi_1 \wedge \pi_2$	$L^* = P_1 \wedge P_2 \wedge e_\infty$
Point pair	$Pp = S_1 \wedge S_2 \wedge S_3$	$Pp^* = P_1 \wedge P_2$

algebra. Geometric objects like spheres, planes, circles, and lines can be represented as vectors and bi-vectors.

Table 4 lists the two representations of the conformal geometric entities. The inner product null space (IPNS) and the outer product null space (OPNS) [61] are dual to each other. While Table 2 already presented the IPNS representation of spheres and planes, they can be described also with the outer product of four points being part of them. In the case of a plane one of these four points is the point at infinity e_∞. Circles can be described with the help of the outer product of three conformal points lying on the circle or as the intersection of two spheres.

Lines can be described with the help of the outer product of two points and the point at infinity e_∞ or with the help of the outer product of two planes (i.e., intersection in IPNS representation). An alternative expression is

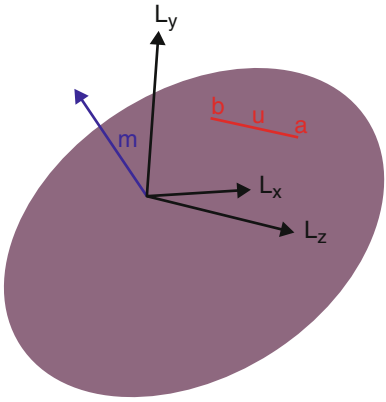


Fig. 20 The line L through the 3D points **a**, **b** and the visualization of its 6D Plücker parameters based on the two 3D vectors **u** and **m** of Eq. (53)

$$L = \mathbf{u}e_{123} + \mathbf{m} \wedge e_{\infty} \tag{52}$$

with the 3D pseudoscalar $e_{123} = e_1 \wedge e_2 \wedge e_{\infty}$, the two 3D points **a**, **b** on the line, $\mathbf{u} = \mathbf{b} - \mathbf{a}$ as 3D direction vector, and $\mathbf{m} = \mathbf{a} \times \mathbf{b}$ as the 3D moment vector (relative to origin). The corresponding six Plücker coordinates (components of **u** and **m**) are (see Fig. 20)

$$(\mathbf{u} : \mathbf{m}) = (u_1 : u_2 : u_3 : m_1 : m_2 : m_3) \tag{53}$$

Computational Efficiency of Geometric Algebra Using Gaalop

Because of its generality, Geometric Algebra needs some optimizations for efficient implementations.

Gaigen [27] is a Geometric Algebra code generator developed at the University of Amsterdam (see [23, 26]). The philosophy behind Gaigen 2 is based on two ideas: generative programming and specializing for the structure of Geometric Algebra. Please find some benchmarks comparing Gaigen 2 with other pure software solutions as well as comparing five models of 3D Euclidean geometry for a ray-tracing application in [26, 28].

Gaalop [44] combines the advantages of software optimizations and the adaptability on different parallel platforms. As an example, an inverse kinematics algorithm of a computer animation application was investigated [42]. With the optimization approach of Gaalop, the software implementation became three times faster and with a hardware implementation about 300 times faster [43] (three times by software optimization and 100 times by additional hardware optimization) than the conventional software implementation. Figure 21 shows an overview over the architecture of Gaalop. Its input is a Geometric Algebra algorithm written in CLUCalc [60], a system for the visual development of Geometric Algebra

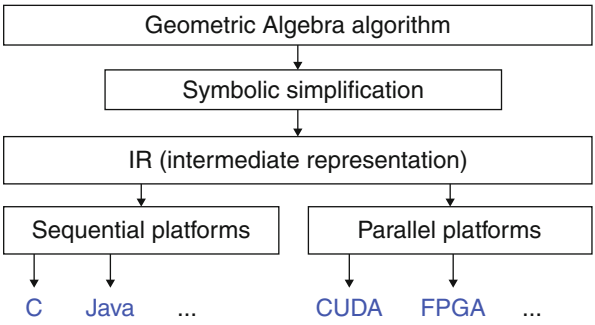


Fig. 21 Architecture of Gaalop

algorithms. Via symbolic simplification it is transformed into an intermediate representation (IR) that can be used for the generation of different output formats. Gaalop supports sequential platforms with the automatic generation of C and JAVA code while its main focus is on supporting parallel platforms like reconfigurable hardware as well as modern accelerating GPUs.

Gaalop uses the symbolic computation functionality of Maple (using the Open Maple interface and a library for Geometric Algebras [1]) in order to optimize a Geometric Algebra algorithm. It computes the coefficients of the desired multivector symbolically, returning an efficient implementation depending just on the input variables.

As an example, the following CLUCalc code computes the intersection circle C of two spheres $S1$ and $S2$ according to Fig. 17:

```
P1 = x1*e1 + x2*e2 + x3*e3 + 1/2*(x1*x1+x2*x2+x3*x3)*einf
     + e0; P2 = y1*e1 +
y2*e2 +y3*e3 + 1/2*(y1*y1+y2*y2+y3*y3)*einf + e0;
     S1 =P1 - 1/2 * r1*r1 *
einf; S2 = P2 - 1/2 * r2*r2 * einf; ?C = S1 $\wedge$ S2;
```

See Table 2 for the computation of the conformal points $P1$ and $P2$, the spheres $S1$ and $S2$, as well as the resulting circle based on the outer product of the two spheres.

The resulting C code generated by Gaalop for the intersection circle C is as follows and depends only on the variables $x1$, $x2$, $x3$, $y1$, $y2$, $y3$, $r1$, and $r2$ for the 3D center points and radii:

```
float C [32] = {\{}0.0{\}}; C[7] = x1*y2-x2*y1;
   C[8] = x1*y3-x3*y1; C[9]
= -0.5*y1*x1*x1-0.5*y1*x2*x2 -0.5*y1*x3*x3+0.5*y1*r1*r1 +
0.5*x1*y1*y1+0.5*x1*y2*y2 + 0.5*x1*y3*y3 - 0.5*x1*r2*r2;
   C[10] = -y1 +
x1; C[11] = -x3*y2+x2*y3; C[12] = -0.5*y2*x1*x1-0.5*y2*x2*x2-
0.5*y2*x3*x3 + 0.5*y2*r1*r1 + 0.5*x2*y1*y1 + 0.5*x2*y2*y2 +
   0.5*x2*y3*y3 -
0.5*x2*r2*r2; C[13] = -y2 + x2; C[14] = -0.5*y3*x1*x1 -
   0.5*y3*x2*x2
```

```
-0.5*y3*x3*x3 + 0.5*y3*r1*r1 + 0.5*x3*y1*y1 + 0.5*x3*y2*y2 +
   0.5*x3*y3*y3
- 0.5*x3*r2*r2; C[15] = -y3 + x3; C[16] = -0.5*y3*y3 +
   0.5*x3*x3 +
0.5*x2*x2 + 0.5*r2*r2 -0.5*y1*y1 - 0.5*y2*y2 + 0.5*x1*x1 -
   0.5*r1*r1;
```

In a nutshell, Gaalop always computes optimized 32-dimensional multivectors. Since a circle is described with the help of a bi-vector, only the blades 7–16 (see Table 3) are used. As you can see, all the corresponding coefficients of this multivector are very simple expressions with basic arithmetic operations.

5 Feature-Based Vector Field Visualization

We will identify derived quantities that describe flow features such as vortices (section "Derived Measures of Vector Fields") and we discuss the topology of vector fields (section "Topology of Vector Fields"). However, not all feature-based visualization approaches can be covered here. The reader is referred to [84] for further information on this topic. We start with a description of integral curves in vector fields, which are the basis for most feature-based visualization approaches.

Characteristic Curves of Vector Fields

A curve $q : \mathbb{R} \rightarrow M$ (see section "Tangential Vectors") is called a *tangent curve* of a vector field $\mathbf{v}(\mathbf{x})$, if for all points $\mathbf{x} \in q$ the tangent vector $\dot{q}$ of q coincides with $\mathbf{v}(\mathbf{x})$. Tangent curves are the solutions of the autonomous ODE system

$$\frac{d}{d\tau}\mathbf{x}(\tau) = \mathbf{v}(\mathbf{x}(\tau)) \quad \text{with } \mathbf{x}(0) = \mathbf{x}_0 \tag{54}$$

For all points $\mathbf{x} \in M$ with $\mathbf{v}(\mathbf{x}) \neq \mathbf{0}$, there is one and only one tangent curve through it. Tangent curves do not intersect or join each other. Hence, tangent curves uniquely describe the directional information and are therefore an important tool for visualizing vector fields.

The tangent curves of a parameter-independent vector field $\mathbf{v}(\mathbf{x})$ are called *streamlines*. A streamline describes the path of a massless particle in $\mathbf{v}$.

In a one-parameter-dependent vector field $\mathbf{v}(\mathbf{x}, t)$, there are four types of characteristic curves: streamlines, path lines, streak lines, and time lines. To ease the explanation, we consider $\mathbf{v}(\mathbf{x}, t)$ as a time-dependent vector field in the following: in a space-time point $(\mathbf{x}_0, t_0)$ we can start a *streamline* (staying in time slice $t = t_0$) by integrating

$$\frac{d}{d\tau}\mathbf{x}(\tau) = \mathbf{v}(\mathbf{x}(\tau), t_0) \quad \text{with } \mathbf{x}(0) = \mathbf{x}_0 \tag{55}$$

or a *path line* by integrating

$$\frac{d}{dt}\mathbf{x}(t) = \mathbf{v}(\mathbf{x}(t), t) \quad \text{with } \mathbf{x}(t_0) = \mathbf{x}_0 \tag{56}$$

Path lines describe the trajectories of massless particles in time-dependent vector fields. The ODE system (56) can be rewritten as an autonomous system at the expense of an increase in dimension by one, if time is included as an explicit state variable:

$$\frac{d}{dt}\begin{pmatrix}\mathbf{x}\\ t\end{pmatrix} = \begin{pmatrix}\mathbf{v}(\mathbf{x}(t), t)\\ 1\end{pmatrix} \quad \text{with } \begin{pmatrix}\mathbf{x}\\ t\end{pmatrix}(0) = \begin{pmatrix}\mathbf{x}_0\\ t_0\end{pmatrix} \tag{57}$$

In this formulation space and time are dealt with on equal footing – facilitating the analysis of spatiotemporal features. Path lines of the original vector field $\mathbf{v}$ in ordinary space now appear as tangent curves of the vector field

$$\mathbf{p}(\mathbf{x}, t) = \begin{pmatrix}\mathbf{v}(\mathbf{x}, t)\\ 1\end{pmatrix} \tag{58}$$

in space-time. To treat streamlines of $\mathbf{v}$, one may simply use

$$\mathbf{s}(\mathbf{x}, t) = \begin{pmatrix}\mathbf{v}(\mathbf{x}, t)\\ 0\end{pmatrix} \tag{59}$$

Figure 22 illustrates $\mathbf{s}$ and $\mathbf{p}$ for a simple example vector field $\mathbf{v}$. It is obtained by a linear interpolation over time of two bilinear vector fields. Figure 22a depicts streamlines, Fig. 22b depicts pathlines.

A *streak line* is the connection of all particles set out at different times but the same point location. In an experiment, one can observe these structures by constantly releasing dye into the flow from a fixed position. The resulting streak line consists of all particles which have been at this fixed position sometime in the past. Considering the vector field $\mathbf{p}$ introduced above, streak lines can be obtained in the following way: apply a stream surface integration in $\mathbf{p}$ where the seeding curve is a straight line segment parallel to the t-axis; a streak line is the intersection of this stream surface with a hyperplane perpendicular to the t-axis (Fig. 22c).

A *time line* is the connection of all particles set out at the same time but different locations, i.e., a line which gets advected by the flow. An analogon in the real world is a yarn or wire thrown into a river, which gets transported and deformed by the flow. However, in contrast to the yarn, a time line can get shorter and longer. It can be obtained by applying a stream surface integration in $\mathbf{p}$ starting at a line with $t = \{\text{const.}\}$ and intersecting it with a hyperplane perpendicular to the t-axis (Fig. 22d).

Streak lines and time lines cannot be described as tangent curves in the spatiotemporal domain. Both types of lines fail to have a property of stream and path lines: they are not locally unique, i.e., for a particular location and time there

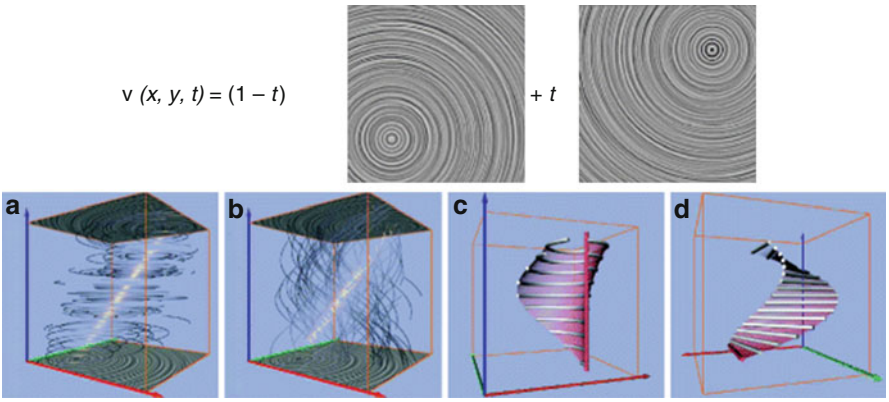


Fig. 22 Characteristic curves of a simple 2D time-dependent vector field. Streamlines and path lines are shown as illuminated field lines. Streak and time lines are shown as thick cylindrical lines, while their seeding curves and resulting stream surfaces are colored *red*. The *red/green* coordinate axes denote the (x, y)-domain; the *blue axis* shows time

is more than one streak and time line passing through. However, stream, path, and streak lines coincide for steady vector fields $\mathbf{v}(\mathbf{x}, t) = \mathbf{v}(\mathbf{x}, t_0)$ and are described by (54) in this setting. Time lines do not fit into this.

Derived Measures of Vector Fields

A number of measures can be derived from a vector field $\mathbf{v}$ and its derivatives. These measures indicate certain properties or features and can be helpful when visualizing flows. The following text assumes the underlying manifold M where the vector field is given to be Euclidean space, i.e., the manifold is three-dimensional and Cartesian coordinates are used where the metric (see section "Tensors") is representable as the unit matrix.

The *magnitude* of $\mathbf{v}$ is then given as

$$|\mathbf{v}| = \sqrt{u^2 + v^2 + w^2} \tag{60}$$

The *divergence* of a flow field is given as

$$\mathrm{div}(\mathbf{v}) = \nabla \cdot \mathbf{v} = \mathrm{trace}(\mathbf{J}) = u_x + v_y + w_z \tag{61}$$

and denotes the gain or loss of mass density at a certain point of the vector field: given a volume element in a flow, a certain amount of mass is entering and exiting it. Divergence is the net flux of this at the limit of a point. A flow field with $\{\mathrm{div}\}(\mathbf{v}) = 0$ is called *divergence-free*, which is a common case in fluid dynamics since a number of fluids are *incompressible*.

The *vorticity* or *curl* of a flow field is given as

$$\omega = \begin{pmatrix} \omega_1 \\ \omega_2 \\ \omega_3 \end{pmatrix} = \nabla \times \mathbf{v} = \begin{pmatrix} w_y - v_z \\ u_z - w_x \\ v_x - u_y \end{pmatrix} \tag{62}$$

This vector is the axis of locally strongest rotation, i.e., it is perpendicular to the plane in which the locally highest amount of circulation takes place. The vorticity magnitude $|\omega|$ gives the strength of rotation and is often used to identify regions of high vortical activity. A vector field with $\omega = \mathbf{0}$ is called *irrotational* or *curl-free*, with the important subclass of *conservative* vector fields, i.e., vector fields which are the gradient of a scalar field. Note that Geometric Algebra (see section "Geometric Algebra" and Sect. 4) treats Eqs. (61) and (62) as an entity, called the geometric derivative. The identification of vortices is a major subject in fluid dynamics. The most widely used quantities for detecting vortices are based on a decomposition of the Jacobian matrix $\mathbf{J} = \mathbf{S} + \Omega$ into its symmetric part, the strain tensor

$$\mathbf{S} = \frac{1}{2}(\mathbf{J} + \mathbf{J}^T) \tag{63}$$

and its antisymmetric part, the vorticity tensor

$$\Omega = \frac{1}{2}(\mathbf{J} - \mathbf{J}^T) = \begin{pmatrix} 0 & -\omega_3 & \omega_2 \\ \omega_3 & 0 & -\omega_1 \\ -\omega_2 & \omega_1 & 0 \end{pmatrix} \tag{64}$$

with ω_i being the components of vorticity (62). While Ω assesses vortical activity, the strain tensor $\mathbf{S}$ measures the amount of stretching and folding which drives mixing to occur.

Inherent to the decomposition of the flow field gradient $\mathbf{J}$ into $\mathbf{S}$ and Ω is the following duality: vortical activity is high in regions where Ω dominates $\mathbf{S}$, whereas strain is characterized by $\mathbf{S}$ dominating Ω.

In order to identify vortical activity, Jeong et al. used this decomposition in [47] to derive the vortex region quantity λ_2 as the second largest eigenvalue of the symmetric tensor $\mathbf{S}^2 + \Omega^2$. Vortex regions are identified by $\lambda_2 < 0$, whereas $\lambda_2 > 0$ lacks physical interpretation. λ_2 does not capture stretching and folding of fluid particles and hence does not capture the vorticity-strain duality.

The Q-criterion of Hunt [46], also known as the Okubo-Weiss criterion, is defined by

$$Q = \frac{1}{2}(\| \Omega \|^2 - \| \mathbf{S} \|^2) = \| \omega \|^2 - \frac{1}{2} \| \mathbf{S} \|^2 \tag{65}$$

where Q is positive and the vorticity magnitude dominates the rate of strain. Hence it is natural to define vortex regions as regions where $Q > 0$. Unlike λ_2, Q has a

physical meaning also where $Q < 0$. Here the rate of strain dominates the vorticity magnitude.

Topology of Vector Fields

In this section we collect the first-order topological properties of steady 2D and 3D vector fields. The extraction of these topological structures has become a standard tool in visualization for the feature-based analysis of vector fields.

Critical Points

Considering a steady vector field $\mathbf{v}(\mathbf{x})$, an isolated *critical point* $\mathbf{x}_0$ is given by

$$\mathbf{v}(\mathbf{x}_0) = \mathbf{0} \quad \text{with} \quad \mathbf{v}(\mathbf{x}_0 \pm \epsilon) \neq \mathbf{0} \tag{66}$$

This means that $\mathbf{v}$ is zero at the critical point but nonzero in a certain neighborhood.

Every critical point can be assigned an *index*. For a 2D vector field it denotes the number of counterclockwise revolutions of the vectors of $\mathbf{v}$ while traveling counterclockwise on a closed curve around the critical point (for 2D vector fields, it is therefore often called the *winding number*). Similarly, the index of a 3D critical point measures the number of times the vectors of $\mathbf{v}$ cover the area of an enclosing sphere. The index is always an integer and it may be positive or negative. For a curve/sphere enclosing an arbitrary part of a vector field, the index of the enclosed area/volume is the sum of the indices of the enclosed critical points. Mann et al. show in [54] how to compute the index of a region using Geometric Algebra. A detailed discussion of index theory can be found in [25, 31, 32].

Critical points are characterized and classified by the behavior of the tangent curves around it. Here we concentrate on first-order critical points, i.e., critical points with $\det(\mathbf{J}(\mathbf{x}_0)) \neq 0$. As shown in [37, 38], a first-order Taylor expansion of the flow around $\mathbf{x}_0$ suffices to completely classify it. This is done by an eigenvalue/eigenvector analysis of $\mathbf{J}(\mathbf{x}_0)$. Let λ_i be the eigenvalues of $\mathbf{J}(\mathbf{x}_0)$ ordered according to their real parts, i.e., $\mathrm{Re}(\lambda_{i-1}) \leq \mathrm{Re}(\lambda_i)$. Furthermore, let $\mathbf{e}_i$ be the corresponding eigenvectors, and let $\mathbf{f}_i$ be the corresponding eigenvectors of the transposed Jacobian $(\mathbf{J}(\mathbf{x}_0))^T$ (note that $\mathbf{J}$ and $\mathbf{J}^T$ have the same eigenvalues but not necessarily the same eigenvectors). The sign of the real part of an eigenvalue λ_i denotes – together with the corresponding eigenvector $\mathbf{e}_i$ – the flow direction: positive values represent an *outflow* and negative values an *inflow* behavior. Based on this we give the classification of 2D and 3D first-order critical points in the following.

2D Vector Fields Based on the flow direction, first-order critical points in 2D vector fields are classified into:

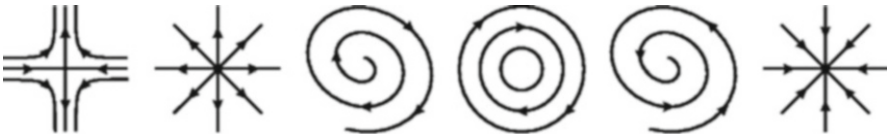


Fig. 23 Classification of first-order critical points. R_1, R_2 denote the real parts of the eigenvalues of the Jacobian matrix, while I_1, I_2 denote their imaginary parts (From [37])

$$\begin{aligned} &\text{Sources :} && 0 < \mathrm{Re}(\lambda_1) \leq \mathrm{Re}(\lambda_2) \\ &\text{Saddles :} && \mathrm{Re}(\lambda_1) < 0 < \mathrm{Re}(\lambda_2) \\ &\text{Sinks :} && \mathrm{Re}(\lambda_1) \leq \mathrm{Re}(\lambda_2) < 0 \end{aligned}$$

Thus, sources and sinks consist of complete outflow/inflow, while saddles have a mixture of both.

Sources and sinks can be further divided into two stable subclasses by deciding whether or not imaginary parts are present, i.e., whether or not λ_1, λ_2 is a pair of conjugate complex eigenvalues:

$$\begin{aligned} &\text{Foci :} && \mathrm{Im}(\lambda_1) = -\mathrm{Im}(\lambda_2) \neq 0 \\ &\text{Nodes :} && \mathrm{Im}(\lambda_1) = \mathrm{Im}(\lambda_2) = 0 \end{aligned}$$

There is another important class of critical points in 2D: a *center*. Here, we have a pair of conjugate complex eigenvalues with $\mathrm{Re}(\lambda_1) = \mathrm{Re}(\lambda_2) = 0$. This type is common in incompressible (divergence-free) flows but unstable in general vector fields since a small perturbation of $\mathbf{v}$ changes the center to either a sink or a source. Figure 23 shows the phase portraits of the different types of first-order critical points following [37].

The index of a saddle point is -1, while the index of a source, sink, or center is $+1$. It turns out that this coincides with the sign of $\det(\mathbf{J}(\mathbf{x}_0))$: a negative determinant denotes a saddle, a positive determinant a source, sink, or center. This already shows that the index of a critical point cannot be used to distinguish or classify them completely, since different types like sources and sinks have assigned the same index.

An iconic representation is an appropriate visualization for critical points, since vector fields usually contain a finite number of them. We will display them as spheres colored according to their classification: sources will be colored in red, sinks in blue, saddles in yellow, and centers in green.

3D Vector Fields Depending on the sign of $\mathrm{Re}(\lambda_i)$ we get the following classification of first-order critical points in 3D vector fields:

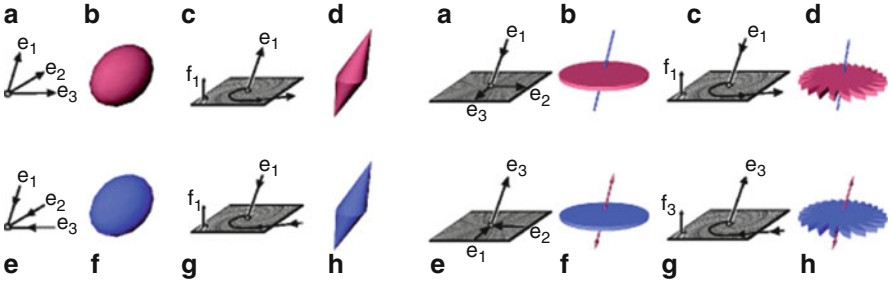


Fig. 24 Flow behavior around critical points of 3D vector fields and corresponding iconic representation

$$\begin{array}{lllll} & \text{Sources}: & & 0 & < \\ \text{Repelling saddles}: & \mathrm{Re}(\lambda_1) < 0 & & < \mathrm{Re}(\lambda_2) \leq & \mathrm{Re}(\lambda_3) \\ \text{Attracting saddles}: & \mathrm{Re}(\lambda_1) \leq & \mathrm{Re}(\lambda_2) < 0 & < & \mathrm{Re}(\lambda_3) \\ & \text{Sinks}: & & \mathrm{Re}(\lambda_1) \leq & \mathrm{Re}(\lambda_2) \leq \end{array}$$

Again, sources and sinks consist of complete outflow/inflow, while saddles have a mixture of both. A repelling saddle has one direction of inflow behavior (called *inflow direction*) and a plane in which a 2D outflow behavior occurs (called *outflow plane*). Similar to this, an attracting saddle consists of an *outflow direction* and an *inflow plane*.

Each of the four classes above can be further divided into two stable subclasses by deciding whether or not imaginary parts in two of the eigenvalues are present ($\lambda_1, \lambda_2, \lambda_3$ are not ordered):

$$\begin{array}{ll} \text{Foci}: & \mathrm{Im}(\lambda_1) = 0 \quad \text{and} \quad \mathrm{Im}(\lambda_2) = -\mathrm{Im}(\lambda_3) \neq 0 \\ \text{Nodes}: & \mathrm{Im}(\lambda_1) = \mathrm{Im}(\lambda_2) = \mathrm{Im}(\lambda_3) = 0 \end{array}$$

As argued in [29], the index of a first-order critical point is given as the sign of the product of the eigenvalues of $\mathbf{J}(\mathbf{x}_0)$. This yields an index of $+1$ for sources and attracting saddles and an index of -1 for sinks and repelling saddles.

In order to depict 3D critical points, several icons have been proposed in the literature; see [30, 36, 37, 53]. Basically, we follow the design approach of [72, 85] and color the icons depending on the flow behavior: attracting parts (inflow) are colored blue, while repelling parts (outflow) are colored red (Fig. 24).

Separatrices

Separatrices are streamlines or stream surfaces which separate regions of different flow behavior. Here we concentrate on separatrices that emanate from critical points. Due to the homogeneous flow behavior around sources and sinks (either a complete outflow or inflow), they do not contribute to separatrices. Each saddle point creates two separatrices: one in forward and one in backward integration into the directions

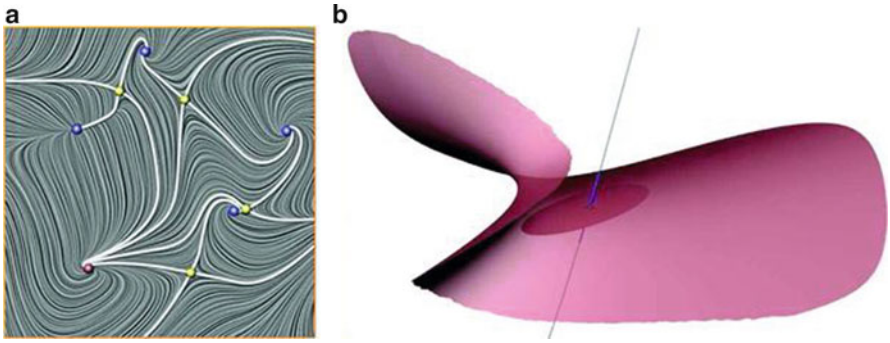


Fig. 25 Separatrices are streamlines or surfaces starting from saddle points into the direction of the eigenvectors

of the eigenvectors. For a 2D saddle point, this gives two separation lines (Fig. 25a). Considering a repelling saddle $\mathbf{x}_R$ of a 3D vector field, it creates one separation curve (which is a streamline starting in $\mathbf{x}_R$ in the inflow direction by backward integration) and a separation surface (which is a stream surface starting in the outflow plane by forward integration). Figure 25b gives an illustration. A similar statement holds for attracting saddles.

Other kinds of separatrices are possible as well: they can emanate from boundary switch curves [85] and attachment and detachment lines [48], or they are closed separatrices without a specific emanating structure [73].

Application

In the following, we exemplify the topological concepts described above by applying them to a 3D vector field. First, we extract the critical points by searching for zeros in the vector field. Based on an eigenvalue/eigenvector analysis, we identify the different types of the critical points. Starting from the saddles, we integrate the separatrices into the directions of the eigenvectors.

Figure 26 visualizes the electrostatic field around a benzene molecule. This data set was calculated on a 101^3 regular grid using the fractional charges method described in [70]. It consists of 184 first-order critical points depicted in Fig. 26a. The separation surfaces shown in Fig. 26b emanate from 78 attracting and 43 repelling saddles. Note how they hide each other as well as the critical points. Even rendering the surfaces in a semitransparent style does not reduce the visual clutter to an acceptable degree. This is one of the major challenges for the topological visualization of 3D vector fields.

Figure 26c shows a possible solution to this problem by showing the 129 *saddle connectors* that we found in this data set. Saddle connectors are the intersection curves of repelling and attracting separation surfaces and have been introduced to the visualization community in [72]. Despite the fact that saddle connectors can only indicate the approximate run of the separation surfaces, the resulting visualization gives more insight into the symmetry and three-dimensionality of the

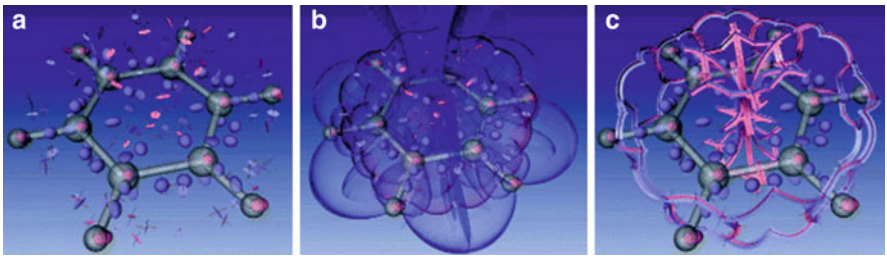


Fig. 26 Topological representations of the benzene data set with 184 critical points. (**a**) Iconic representation. (**b**) Due to the shown separation surfaces, the topological skeleton of the vector field looks visually cluttered. (**c**) Visualization of the topological skeleton using saddle connectors

data set. Saddle connectors are a useful compromise between the amount of coded information and the expressiveness of the visualization for complex topological skeletons.

6 Anisotropic Diffusion PDEs for Image Regularization and Visualization

Regularization PDEs: A Review

We consider a 2D multivalued image $\mathbf{I}$: $\Omega \rightarrow \mathbb{R}^n (n = 3$ for color images) defined on a domain $\Omega \subset \mathbb{R}^2$ and denote by $I_i : \Omega \rightarrow \mathbb{R}$ the scalar channel i of $\mathbf{I}$:

$$\forall \mathbf{X} = (x, y) \in \Omega, \mathbf{I}_{(\mathbf{X})} = (I_{1(\mathbf{X})} \;\; I_{2(\mathbf{X})} \;\; \ldots \;\; I_{n(\mathbf{X})})^T.$$

Local Multivalued Geometry and Diffusion Tensors

PDE-based regularization can be often seen as the local smoothing of an image $\mathbf{I}$ along defined directions depending themselves on the local configuration of the pixel intensities, i.e., one wants basically to smooth $\mathbf{I}$ in parallel to the image discontinuities. Naturally, this means that one has first to retrieve the *local geometry* of the image $\mathbf{I}$. It consists in the definition of these important features at each image point $\mathbf{X} = (x, y) \in \Omega$:

- Two orthogonal directions ${\theta_{(\mathbf{X})}}^+, \; {\theta_{(\mathbf{X})}}^- \in \mathrm{S}^1$ along the local maximum and minimum variations of image intensities at $\mathbf{X}$. θ^- is then considered to be parallel to the local edge, when there is one.
- Two corresponding positive values ${\lambda_{(\mathbf{x})}}^+, {\lambda_{(\mathbf{x})}}^-$ measuring the effective variations of the image intensities along ${\theta_{(\mathbf{x})}}^+$ and ${\theta_{(\mathbf{x})}}^-$, respectively. λ^-, λ^+ are related to the local *contrast* of an edge.

For scalar images $I : \Omega \rightarrow \mathbb{R}$, this local geometry $\{\lambda^{+/-}, \theta^{+/-} | \mathbf{X} \in \Omega\}$ is usually retrieved by the computation of the smoothed gradient field $\nabla \; I_\sigma =$

$\nabla I * G_\sigma$ where G_σ is a 2D Gaussian kernel with standard deviation σ. Then, $\lambda^+ = \| \nabla I_\sigma \|^2$ is a possible measure of the local contrast of the contours, while $\theta^- = \nabla I_\sigma{}^\perp / \| \nabla I_\sigma \|$ gives the contours direction. Such a local geometry $\{\lambda^{+/-}, \theta^{+/-} | \mathbf{X} \in \Omega\}$ can be represented in a more convenient form by a field $\mathbf{G}$: $\Omega \rightarrow$ P(2) of second-order tensors (2×2 symmetric and semi-positive matrices):

$$\forall \mathbf{X} \in \Omega, \quad \mathbf{G_{(X)}} = \lambda^- \, \theta^- \theta^{-^T} + \lambda^+ \, \theta^+ \theta^{+^T}.$$

Eigenvalues of $\mathbf{G}$ are indeed λ^- and λ^+ and corresponding eigenvectors are θ^- and θ^+. The local geometry of scalar-valued images I can be then modeled by the tensor field $\mathbf{G_{(x)}} = \nabla I_{\sigma(\mathbf{x})} \nabla I^T_{\sigma(\mathbf{x})}$.

For multivalued images $\mathbf{I} : \Omega \rightarrow \mathbb{R}^n$, the local geometry can be retrieved in a similar way, by the computation of the field $\mathbf{G}$ of the smoothed *structure tensors*. As explained in [22,82], this is a nice extension of the gradient for multivalued images:

$$\forall \mathbf{X} \in \Omega, \quad \mathbf{G}_{\sigma(\mathbf{X})} = \left(\sum\nolimits_{i=1}^{n} \nabla I_{i\alpha(\mathbf{X})} \nabla I^T_{i\alpha(\mathbf{X})} \right)_* G_\sigma \quad \text{where} \quad \nabla I_{i\alpha} = \begin{pmatrix} \frac{\partial I_i}{\partial x} \\ \frac{\partial I_i}{\partial y} \end{pmatrix} * G_\alpha \tag{67}$$

$\mathbf{G}_{\sigma(\mathbf{x})}$ is a very good estimator of the local multivalued geometry of $\mathbf{I}$ at $\mathbf{X}$: its spectral elements give at the same time the vector-valued variations (by the eigenvalues λ^-, λ^+ of $\mathbf{G}_\sigma$) and the orientations (edges) of the local image structures (by the eigenvectors $\theta^- \perp \theta^+$ of $\mathbf{G}_\sigma$), σ being proportional to the so-called noise scale.

Once the local geometry $\mathbf{G}_\sigma$ of $\mathbf{I}$ has been determined, the way the regularization process is achieved is defined by another field $\mathbf{T}$: $\Omega \rightarrow$ P(2) of *diffusion tensors*, which specifies the local smoothing geometry that should drive the PDE flow. Of course, $\mathbf{T}$ depends on the targeted application, and most of the time it is constructed from the local geometry $\mathbf{G}_\sigma$ of $\mathbf{I}$. It is thus defined from the spectral elements λ^-, λ^+ and θ^-, θ^+ of $\mathbf{G}_\sigma$. In [19, 78], the following expression is proposed for image regularization:

$$\forall \mathbf{X} \in \Omega, \quad \mathbf{T_{(X)}} = f^-_{(\lambda^+, \lambda^-)} \, \theta^- \theta^- + f^+_{(\lambda^+, \lambda^-)} \, \theta^+ \theta^{+^T} \tag{68}$$

where

$$f^-_{(\lambda_+, \lambda_-)} = \frac{1}{(1 + \lambda^+ + \lambda^-)^{p_1}} \quad \text{and} \quad f^+_{(\lambda_+, \lambda_-)} = \frac{1}{(1 + \lambda^+ + \lambda^-)^{p_2}} \quad \text{with } p_1 < p_2$$

are the two functions which set the strengths of the desired smoothing along the respective directions θ^-, θ^+. This latest choice basically says that if a pixel $\mathbf{X}$ is located on an image contour ($\lambda^+_{(\mathbf{x})}$ is high), the smoothing on $\mathbf{X}$ would be performed mostly along the contour direction $\theta^-_{(\mathbf{x})}$ (since $f^+_{(...)} \ll f^-_{(...)}$). Conversely, if a pixel $\mathbf{X}$ is located on a homogeneous region ($\lambda^+_{(\mathbf{x})}$ is low), the smoothing on $\mathbf{X}$ would be

performed in all possible directions (isotropic smoothing), since $f^+_{(...)} \simeq f^-_{(...)}$ (and then $\mathbf{T} \simeq \mathbb{I}_d$). Predefining the smoothing geometry $\mathbf{T}$ of each applied PDE iteration is the first stage of most of the PDE-based regularization algorithms. Most of the differences between existing regularization methods (as in [2,3,10,18,19,49,52,58,59,67–69]) lie first on the definition of $\mathbf{T}$, but also on the kind of the diffusion PDE that will be used indeed to perform the desired smoothing.

Divergence-Based PDEs

One of the common choices to smooth a corrupted multivalued image $\mathbf{I}:\Omega \to \mathbb{R}^n$ following a local smoothing geometry $\mathbf{T}$: $\Omega \to \mathrm{P}(2)$ is to use the divergence PDE:

$$\forall i = 1, \ldots, n, \quad \frac{\partial I_i}{\partial t} = \mathrm{div}(\mathbf{T}\nabla I_i) \tag{69}$$

The general form of this now classical PDE for image regularization has been introduced by Weickert in [82] and adapted for color/multivalued images in [83]. In this latter case, the tensor field $\mathbf{T}$ is chosen the same for all image channels I_i, ensuring that channels are smoothed with a *coherent multivalued geometry* which takes the correlation between channels into account (since $\mathbf{T}$ depends on $\mathbf{G}$). Equation (69) unifies a lot of existing scalar or multivalued regularization approaches and proposes at the same time two interpretation levels of the regularization process:

- *Local interpretation*: Equation (69) may be seen as the physical law describing local diffusion processes of the pixels individually regarded as temperatures or chemical concentrations in an anisotropic environment which is locally described by $\mathbf{T}$.
- *Global interpretation*: The problem of image regularization can be regarded as the minimization of the energy functional $E(\mathbf{I})$ by a gradient descent (i.e., a PDE), coming from the Euler-Lagrange equations of $E(\mathbf{I})$ [3,19,49,51,78]:

$$E(\mathbf{I}) = \int_\Omega \psi(\lambda^+, \lambda^-)\, d\Omega \qquad \text{where } \psi : \mathbb{R}^2 \to \mathbb{R} \tag{70}$$

- It results in a particular case of the PDE (69), with $\mathbf{T} = \frac{\partial \Psi}{\partial \lambda^-}\theta^-\theta^{-^T} + \frac{\partial \Psi}{\partial \lambda^+}\theta^+\theta^{+^T}$, where λ_+, λ_- are the two positive eigenvalues of the *non-smoothed* structure tensor field $\mathbf{G} = \sum_i \nabla I_i\ \nabla I_i^T$ and θ_+, θ_- are the corresponding eigenvectors.

Unfortunately, there are local configurations where the PDE (69) *does not fully respect the geometry* $\mathbf{T}$ and where the smoothing is performed in unexpected directions. For instance, considering (69) with tensor fields $\mathbf{T}_{1(\mathbf{X})} = \left(\frac{\nabla I}{\|\nabla I\|}\right)\left(\frac{\nabla I}{\|\nabla I\|}\right)^T$ (purely anisotropic) and $\mathbf{T}_{2(x)} = \mathbb{I}_d$ (purely isotropic) lead both to the heat equation $\frac{\partial I}{\partial t} = \Delta I$ which has obviously an isotropic smoothing behavior. Different tensors fields $\mathbf{T}$ with different shapes (isotropic or anisotropic) may define the same regularization behavior. This is due to the fact that the divergence implicitly

introduces a dependance on the *spatial variations* of the tensor field $\mathbf{T}$, so it hampers the design of a pointwise smoothing behavior.

Trace-Based PDEs

Alternative PDE-based regularization approaches have been proposed in [3, 51, 68, 69, 78] in order to smooth an image directed by a local smoothing geometry. They are inspirit very similar to the divergence equation (69), but based on a *trace* operator:

$$\forall i = 1, \ldots, n, \quad \frac{\partial I_i}{\partial t} = \text{trace}\,(\mathbf{T}\mathbf{H}_i) \qquad \text{with } \mathbf{H}_i = \begin{pmatrix} \frac{\partial^2 I_i}{\partial x^2} & \frac{\partial^2 I_i}{\partial x \partial y} \\ \frac{\partial^2 I_i}{\partial x \partial y} & \frac{\partial^2 I_i}{\partial y^2} \end{pmatrix} \tag{71}$$

$\mathbf{H}_i$ stands for the Hessian of I_i. Equation (71) is in fact nothing more than a tensor-based expression of the PDE $\frac{\partial \mathbf{I}}{\partial t} = f^-_{(\lambda^-,\lambda^+)}\, \mathbf{I}_{\theta^-\theta^-} + f^+_{(\lambda^-,\lambda^+)}\, \mathbf{I}_{\theta^+\theta^+}$ where $\mathbf{I}_{\theta^-\theta^-} = \frac{\partial^2 \mathbf{I}}{\partial \theta^{-2}}$. This PDE can be viewed as a simultaneous combination of two orthogonally oriented and weighted 1D Laplacians. In case of multivalued images, each channel I_i of $\mathbf{I}$ is here also coherently smoothed with the same tensor field $\mathbf{T}$. As demonstrated in [78], the evolution of Eq. (71) has a geometric meaning in terms of local linear filtering: it may be seen locally as the application of very small convolutions around each point $\mathbf{X}$ with a Gaussian mask $G_t^{\mathbf{T}}$ *oriented* by the tensor $\mathbf{T}_{(x)}$:

$$G^{\mathbf{T}}_{t(\mathbf{X})} = \frac{1}{4\pi t} \exp\left(-\frac{\mathbf{X}^T\, \mathbf{T}^{-1}\, \mathbf{X}}{4t}\right)$$

This ensures that the smoothing performed by (71) is indeed oriented along the predefined smoothing geometry $\mathbf{T}$. As the trace is not a differential operator, the spatial variation of $\mathbf{T}$ does not trouble the diffusion directions here and two different tensor fields will necessarily lead to different smoothing behaviors. Under certain conditions, the divergence PDE (69) may be also developed as a trace formulation (71). But in this case, the tensors inside the trace and the divergence *are not the same* [78]. Note that trace-based equations (71) are rather directly connected to functional minimizations, especially when considering the multivalued case. For scalar-valued images ($n = 1$), some correspondences are known anyway [3, 19, 51].

Curvature-Preserving PDEs

Basically, the divergence and trace Eqs. (69) and (71) locally behave as oriented Gaussian smoothing whose strengths and orientations are directly related to the tensors $\mathbf{T}_{(x)}$. But on curved structures (like corners), this behavior is not desirable: in case of high variations of the edge orientation θ^-, such a smoothing will tend to *round* corners, even by conducting it only along θ^- (an oriented Gaussian is not curved by itself). To avoid this over-smoothing effect, regularization PDEs may try to stop their action on corners (by vanishing tensors $\mathbf{T}_{(x)}$ there, i.e., $f^- = f^+ = 0$),

but this implies the detection of curved structures on images that are themselves noisy or corrupted. This is generally a hard task.

To overcome this problem, curvature-preserving regularization PDEs have been introduced in [77]. We illustrate the general idea of these equations by considering the simplest case of image smoothing along a single direction, i.e., a *vector field* $\mathbf{w} : \Omega \to \mathbb{R}^2$ instead of a tensor-valued one $\mathbf{T}$. The two spatial components of $\mathbf{w}$ are denoted $\mathbf{w}_{(x)} = (u_{(x)} v_{(x)})^T$.

The curvature-preserving regularization PDE that smoothes $\mathbf{I}$ along $\mathbf{w}$ is defined as

$$\forall i = 1, \ldots, n, \quad \frac{\partial I_i}{\partial t} = \text{trace}\left(\mathbf{w}\mathbf{w}^T \mathbf{H}_i\right) + \nabla I_i^T \mathbf{J}_\mathbf{w} \mathbf{w} \qquad \text{with } \mathbf{J}_\mathbf{w} = \begin{pmatrix} \frac{\partial u}{\partial x} & \frac{\partial u}{\partial y} \\ \frac{\partial v}{\partial x} & \frac{\partial v}{\partial y} \end{pmatrix} \tag{72}$$

where $\mathbf{J}_\mathbf{w}$ stands for the Jacobian of $\mathbf{w}$. Equation (72) simply adds a term $\nabla I_i^T \mathbf{J}_\mathbf{w} \mathbf{w}$ to the corresponding trace-based PDE (71) that would smooth $\mathbf{I}$ along $\mathbf{w}$. This term naturally depends on the variation of the vector field $\mathbf{w}$. Actually, it has been demonstrated in [77] that Eq. (72) is equivalent to the application of this one-dimensional PDE flow:

$$\frac{\partial I_i(\mathcal{C}_{(a)})}{\partial t} = \frac{\partial^2 I_i(\mathcal{C}_{(a)})}{\partial a^2} \qquad \text{with } \begin{cases} \mathcal{C}^\mathbf{X}_{(0)} = \mathbf{X} \\ \frac{\partial \mathcal{C}^\mathbf{X}_{(a)}}{\partial a} = \mathbf{w}\left(\mathcal{C}^\mathbf{X}_{(a)}\right) \end{cases} \tag{73}$$

where $\mathcal{C}^\mathbf{X}_{(a)}$ is the streamline curve of $\mathbf{w}$, starting from $\mathbf{X}$ and parameterized by $a \in \mathbb{R}$. Thus, Eq. (73) is nothing more than the *one-dimensional heat flow constrained on the streamline curve* $\mathcal{C}$. This is indeed very different from a heat flow *oriented* by $\mathbf{w}$, as in the formulation $\frac{\partial I_i}{\partial t} = \frac{\partial^2 I_i}{\partial \mathbf{w}^2}$ since the curvatures of the streamline of $\mathbf{w}$ are now implicitly taken into account. In particular, Eq. (73) has the interesting property to vanish when the image intensities are constant on the streamline $\mathcal{C}^\mathbf{X}$, whatever the curvature of $\mathcal{C}^\mathbf{X}$ is. So, defining a field $\mathbf{w}$ that is tangent everywhere to the image structures allows the preservation of these structures during the regularization process, even if they are curved (such as corners).

Moreover, as Eq. (73) is a 1D heat flow on a streamline $\mathcal{C}^\mathbf{X}$, its solution at time dt can be estimated by convolving the image signal lying on the streamline $\mathcal{C}^\mathbf{X}$ by a 1D Gaussian kernel [50]:

$$\forall \mathbf{X} \in \Omega, \qquad \mathbf{I}^{[dt]}_{(\mathbf{X})} = \int_{-\infty}^{+\infty} \mathbf{I}^{[t=0]}\left(\mathcal{C}^\mathbf{X}_{(p)}\right) \; G^{dt}(p) \; dp \tag{74}$$

This formulation is very close to the line integral convolution (LIC) framework [17], which has been introduced as a visualization technique to render a textured image representing a 2D vector field $\mathbf{w}$. As we are considering diffusion equations here, the weighting function in Eq. (74) is naturally Gaussian. This geometric interpretation particularly allows to implement curvature-preserving PDEs (74) using Runge-

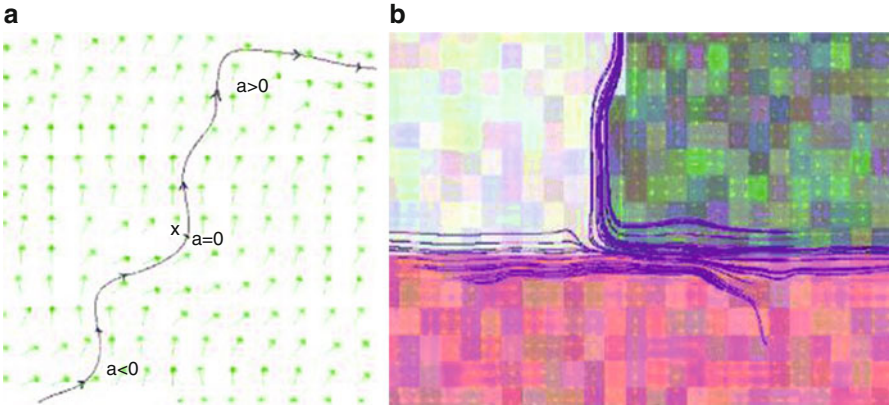


Fig. 27 Streamline $\mathcal{C}^X$ of various vector fields $\mathbf{w}$:$\Omega \rightarrow \mathbb{R}^2$. (**a**) Streamline of a general field $\mathbf{w}$. (**b**) Example of streamlines when $\mathbf{w}$ is the lowest eigenvector of the smoothed structure tensor $\mathbf{G}_\sigma$ (one block is one color pixel)

Kutta estimations of the streamline geometries, leading to sub-pixel precision of the smoothing process.

This single-direction smoothing PDE (72) can be easily extended to deal with a tensor-valued geometry $\mathbf{T}$: $\Omega \rightarrow$ P(2), in order to be able to represent both *anisotropic* or *isotropic* regularization behaviors. This is done by decomposing the tensor field $\mathbf{T}$ as the sum of several single-directional tensors, i.e., $\mathbf{T} = \frac{2}{\pi}\int_{\alpha=0}^{\pi}\left(\sqrt{\mathbf{T}}a_\alpha\right)\left(\sqrt{\mathbf{T}}a_\alpha\right)^T d\alpha$, where $a_\alpha = \cos\alpha\ \sin\alpha^T$. This naturally suggests to decompose a tensor-driven regularization process into a sum of single-direction smoothing processes, each of them being expressed as a curvature-preserving PDE. As a result, the corresponding curvature-preserving PDE directed by a tensor field $\mathbf{T}$ is

$$\forall i = 1,\ldots,n, \qquad \frac{\partial I_i}{\partial t} = \mathrm{trace}(\mathbf{T}\mathbf{H}_i) + \frac{2}{\pi}\nabla I_i^T \int_{\alpha=0}^{\pi} \mathbf{J}_{\sqrt{\mathbf{T}}a_a}\sqrt{\mathbf{T}}a_a\, d\alpha \qquad (75)$$

When $\mathbf{T}$ is locally isotropic (on homogeneous region), Eq. (75) is similar to a 2D heat flow, while when $\mathbf{T}$ is locally anistropic (on an image contour), it behaves as a 1D heat flow on the streamline curve following the contour path, thus taking care of its curvature (Fig. 27).

Applications

Some application results are presented here, mainly based on the use of the curvature-preserving PDEs (75). A specific diffusion tensor field $\mathbf{T}$ has been used to adapt the smoothing behavior to each targeted application.

Noisy color image (*left*), denoised image (*right*) by curvature-preserving PDE (50.75)

Image of a fingerprint

After several iterations of trace-based PDE (50.71)

After several iterations of curvature-preserving PDE (50.75) (with same tensor field T)

Fig. 28 Using PDE-based smoothing to regularize color and grayscale images

Original color image (*left*), image with 50% pixel removed (*middle*), reconstructed using PDE (50.75) (*right*)

Original color image (*left*), reconstructed using PDE (50.75) (*right*) (the in painting mask covers the cage)

Fig. 29 Image inpainting using PDE-based regularization techniques

Color Image Denoising

Image denoising is a direct application of regularization methods. Sensor inaccuracies, digital quantifications, or compression artifacts are indeed some of the various noise sources that can affect a digital image, and suppressing them is a desirable goal. Figure 28 illustrates how curvature-preserving PDEs (75) can be successfully applied to remove such noise artifacts while preserving the thin structures of the processed images. The tensor field $\mathbf{T}$ is chosen as in Eq. (68).

Color Image Inpainting

Image inpainting consists in filling in missing (user-defined) image regions by guessing pixel values such that the reconstructed image still looks natural. Basically, the user provides one color image $\mathbf{I} : \Omega \to \mathbb{R}^3$ and one *mask* image $M : \Omega \to \{0, 1\}$.

The inpainting algorithm must fill in the regions where $M(X) = 1$, by means of some intelligent interpolations. Image inpainting using diffusion PDEs has been proposed, for instance, in [10, 18, 78]. Inpainting is a direct application of our proposed curvature-preserving PDE (75), where the diffusion equation is applied only on the regions to inpaint, allowing the neighbor pixels to diffuse inside these regions in an anisotropic way (Fig. 29).

Visualization of Vector and Tensor Fields

Regularization PDEs such as (69), (71), and (75) can be also used to visualize a vector field $\mathbf{w} : \Omega \to \mathbb{R}^2$ or a tensor field $\mathbf{G}$: $\Omega \to \mathrm{P}(2)$; see also Sect. 5. The idea is to smooth an originally pure noisy image using a diffusion tensor field $\mathbf{T}$ which is chosen to be $\mathbf{T} = \mathbf{w}\mathbf{w}^T$ or $\mathbf{T} = \mathbf{G}$ or other variations as long as the smoothing geometry is indeed directed by the field we want to visualize. Whereas the PDE evolution time t goes by, more global structures of the considered fields appear, i.e., a visualization *scale-space* is constructed. The same PDE-based visualization technique allows to display interesting global rendering of DT-MRI volumes (medical imaging) displaying "stuffed" views of the fiber map (Fig. 30).

7 Conclusion

This chapter presented a selection of possible approaches for systematic treatment of multidimensional data sets and algorithms based on differential methods. Such data sets may be the result of some image processing, for instance, a three-dimensional stack of CT slices in medical imaging used to extract a triangular surface representing bones. Visual data analysis is particularly important for big data. Many such data sources originate in computational sciences, produced by algorithms based on differential methods. Utilizing the same concepts for image processing and multidimensional data in general allows generalization of methods and increased software reusability and applicability. Section 2 reviews a mathematically founded model to structure multidimensional data covering a wide category of data types. Since there is no commonly agreed standard in the scientific community on how to lay out multidimensional data for computational purposes, a multitude of alternative models exist. For a specific application, it is subject of investigation whether some specific model would fit all the respective requirements. Section 3 delves deeper into the modeling of mathematical operators using computational data structures. A mathematical algorithm must be cast into a discretized form in order to be applicable in a numerical code. Differential forms are a mathematical abstraction allowing coordinate-free formulations of partial differential equations, which are the basis of many physical and engineering methods, as well as image processing. Geometric Algebra, reviewed in Sect. 4, extends the notions of the commonly known vector algebra to form a full system of algebraic operations to include (among others) the notion of "dividing vectors." While unfamiliar at first, the result is a visually intuitive way to phrase complex algebraic operations, enabling better insight and more efficient implementation as compared to "ad hoc" approaches. An important

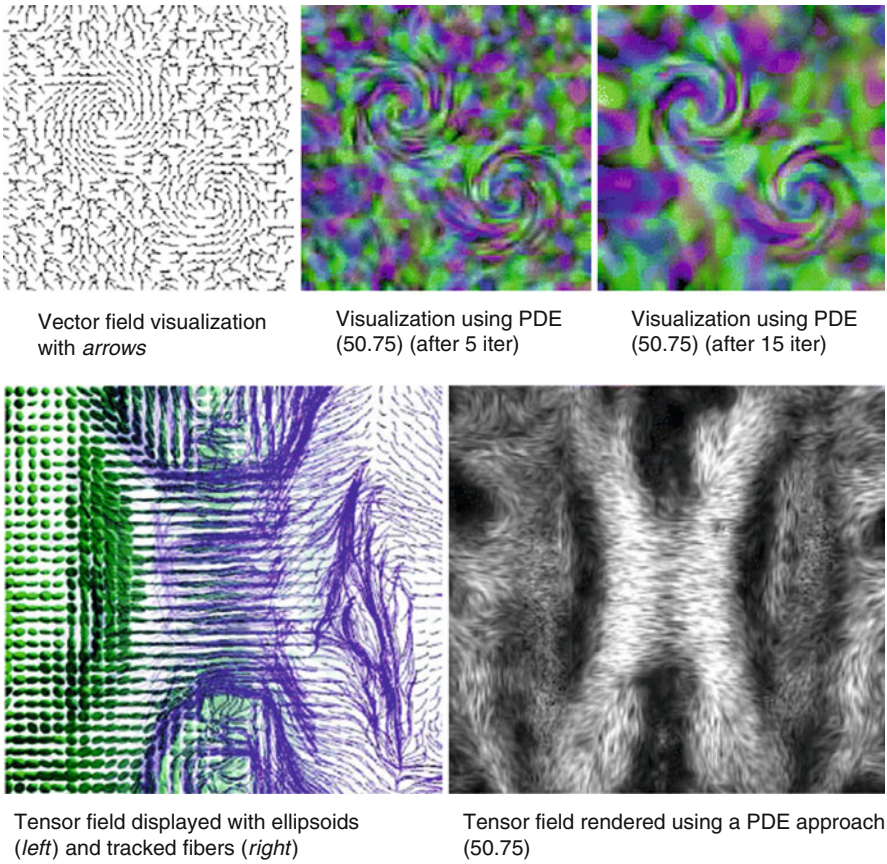

Vector field visualization with *arrows*
Visualization using PDE (50.75) (after 5 iter)
Visualization using PDE (50.75) (after 15 iter)
Tensor field displayed with ellipsoids (*left*) and tracked fibers (*right*)
Tensor field rendered using a PDE approach (50.75)

Fig. 30 Visualization of vector and tensor fields using PDEs

application of differential methods for multidimensional data is the investigation of features in vector fields. While primarily of interest to computational sciences such as computational fluid dynamics, identifying topological features in vector fields may well apply to color images with RGB channels and gradients of grayscale images or even more to stacks of images such as three-dimensional data and animation sequences. Section 5 presents some basics for feature detections in such data sets. While the application to animation sequences is beyond the scope of this chapter, utilizing the topological skeleton of some data set may well have potential for data compression or automatizing algorithms for motion pictures. Partial differential equations (PDEs) furthermore allow for direct improvement of image quality and feature reconstruction, as explained in the algorithms presented in Sect. 6. PDEs describing diffusion are particularly suitable for reducing noise in images, recovering lost features, as well as the direct visualization of vector and tensor fields. The set of presented algorithms in this chapter is still subject of active

research and neither a comprehensive nor the only reasonable approach; rather, the various aspects covered provide inspirations covering many scientific disciplines under one hood.

Cross-References

- ▸ Tomography
- ▸ Mathematical Methods in PET and SPECT Imaging
- ▸ Mathematics of Electron Tomography

References

1. Abłamowicz, R., Fauser, B.: Clifford/bigebra, a maple package for Clifford (co)algebra computations. Available at http://www.math.tntech.edu/rafal/. © 1996–2009, RA&BF (2009)
2. Alvarez, L., Guichard, F., Lions, P.L., Morel, J.M.: Axioms and fundamental equations of image processing. Arch. Ration. Mech. Anal. **123**(3), 199–257 (1993)
3. Aubert, G., Kornprobst, P.: Mathematical Problems in Image Processing: Partial Differential Equations and the Calculus of Variations. Applied Mathematical Sciences, vol. 147. Springer, New York (2002)
4. Barash, D.: A fundamental relationship between bilateral filtering, adaptive smoothing and the nonlinear diffusion equation. IEEE Trans. Pattern Anal. Mach. Intell. **24**(6), 844 (2002)
5. Bayro-Corrochano, E., Vallejo, R., Arana-Daniel, N.: Geometric preprocessing, geometric feedforward neural networks and Clifford support vector machines for visual learning. Spec. Issue J. Neurocomput. **67**, 54–105 (2005)
6. Becker, J., Preusser, T., Rumpf, M.: PDE methods in flow simulation post processing. Comput. Vis. Sci. **3**(3), 159–167 (2000)
7. Benger, W.: Visualization of general relativistic tensor fields via a fiber bundle data model. PhD thesis, FU Berlin (2004)
8. Benger, W.: Colliding galaxies, rotating neutron stars and merging black holes – visualising high dimensional data sets on arbitrary meshes. N. J. Phys. **10** (2008). http://stacks.iop.org/1367-2630/10/125004
9. Benger, W., Ritter, M, Acharya, S., Roy, S., Jijao, F.: Fiberbundle-based visualization of a stir tank fluid. In: WSCG 2009, Plzen (2009)
10. Bertalmio, M., Sapiro, G., Caselles, V., Ballester, C.: Image inpainting. In: ACM SIGGRAPH, International Conference on Graphics and Interactive Techniques, New Orleans, pp. 417–424 (2000)
11. Black, M.J., Sapiro, G., Marimont, D.H., Heeger, D.: Robust anisotropic diffusion. IEEE Trans. Image Process. **7**(3), 421–432 (1998)
12. Bochev, P., Hyman, M.: Principles of compatible discretizations. In: The IMA Hot Topics Workshop on Compatible Discretizations, University of Minnesota, 11–15 May 2004. IMA, vol. 142, pp. 89–120. Springer (2006)
13. Brouwer, L.: Zur Invarianz des n-dimensionalen Gebiets. Math. Ann. **71**, 305–313 (1912)
14. Buchholz, S., Hitzer, E.M.S., Tachibana, K.: Optimal learning rates for Clifford neurons. In: International Conference on Artificial Neural Networks, Porto, vol. 1, pp. 864–873, 9–13 (2007)
15. Butler, D.M., Bryson, S.: Vector bundle classes form a powerful tool for scientific visualization. Comput. Phys. **6**, 576–584 (1992)
16. Butler, D.M., Pendley, M.H.: A visualization model based on the mathematics of fiber bundles. Comput. Phys. **3**(5), 45–51 (1989)

17. Cabral, B., Leedom, L.C.: Imaging vector fields using line integral convolution. In: SIGGRAPH'93, in Computer Graphics, Anaheim, vol. 27, pp. 263–272 (1993)
18. Chan, T., Shen, J.: Non-texture inpaintings by curvature-driven diffusions. J. Vis. Commun. Image Represent. **12**(4), 436–449 (2001)
19. Charbonnier, P., Blanc-Féraud, L., Aubert, G., Barlaud, M.: Deterministic edge-preserving regularization in computed imaging. IEEE Trans. Image Process. **6**(2), 298–311 (1997)
20. Clifford, W.K.: Applications of Grassmann's extensive algebra. In: Tucker, R. (ed.) Mathematical Papers, pp. 266–276. Macmillian, London (1882)
21. Clifford, W.K.: On the classification of geometric algebras. In: Tucker, R. (ed) Mathematical Papers, pp. 397–401. Macmillian, London (1882)
22. Di Zenzo, S.: A note on the gradient of a multi-image. Comput. Vis. Graph. Image Process. **33**, 116–125 (1986)
23. Dorst, L., Fontijne, D., Mann, S.: Geometric Algebra for Computer Science, an Object-Oriented Approach to Geometry. Morgan Kaufman, San Mateo (2007)
24. Ebling, J.: Clifford Fourier transform on vector fields. IEEE Trans. Vis. Comput. Graph. **11**(4), 469–479. IEEE member Scheuermann, Gerik (2005)
25. Firby, P., Gardiner, C.: Surface Topology, chap. 7, pp. 115–135. Ellis Horwood, Chichester (1982). Vector Fields on Surfaces
26. Fontijne, D.: Efficient implementation of geometric algebra. PhD thesis, University of Amsterdam (2007)
27. Fontijne, D., Bouma, T., Dorst, L.: Gaigen: a geometric algebra implementation generator. http://www.science.uva.nl/ga/gaigen (2005)
28. Fontijne, D., Dorst, L.: Modeling 3D Euclidean geometry. IEEE Comput. Graph. Appl. **23**(2), 68–78 (2003)
29. Garth, C., Tricoche, X., Scheuermann, G.: Tracking of vector field singularities in unstructured 3D time-dependent datasets. In: Proceedings of the IEEE Visualization, Austin, pp. 329–336 (2004)
30. Globus, A., Levit, C., Lasinski, T.: A tool for visualizing the topology of threedimensional vector fields. In: Proceedings of the IEEE Visualization '91, San Diego, pp. 33–40 (1991)
31. Gottlieb, D.H.: Vector fields and classical theorems of topology. Rend. Semin. Mat. Fisico, Milano **60**, 193–203 (1990)
32. Gottlieb, D.H.: All the way with Gauss-Bonnet and the sociology of mathematics. Am. Math. Mon. **103**(6), 457–469 (1996)
33. Gross, P., Kotiuga, P.R.: Electromagnetic Theory and Computation: A Topological Approach. Cambridge University Press, Cambridge (2004)
34. Hart, J.: Using the CW-complex to represent the topological structure of implicit surfaces and solids. In: Implicit Surfaces '99, Eurographics/SIGGRAPH, Bordeaux, 13–15 Sept 1999, pp. 107–112
35. Hatcher, A.: Algebraic Topology. Cambridge University Press, Cambridge (2002)
36. Hauser, H., Gröller, E.: Thorough insights by enhanced visualization of flow topology. In: 9th International Symposium on Flow Visualization, Edinburgh (2000). http://www.cg.tuwien.ac.at/research/publications/2000/Hauser-2000-Tho/
37. Helman, J., Hesselink, L.: Representation and display of vector field topology in fluid flow data sets. IEEE Comput. **22**(8), 27–36 (1989)
38. Helman, J., Hesselink, L.: Visualizing vector field topology in fluid flows. IEEE Comput. Graph. Appl. **11**, 36–46 (1991)
39. Hestenes, D.: New Foundations for Classical Mechanics. Reidel, Dordrecht (1986)
40. Hestenes, D., Sobczyk, G.: Clifford Algebra to Geometric Calculus: A Unified Language for Mathematics and Physics. Reidel, Dordrecht (1984)
41. Hildenbrand, D., Fontijne, D., Perwass, C., Dorst, L.: Tutorial geometric algebra and its application to computer graphics. In: Eurographics Conference, Grenoble (2004)
42. Hildenbrand, D., Fontijne, D., Wang, Y., Alexa, M., Dorst, L.: Competitive runtime performance for inverse kinematics algorithms using conformal geometric algebra. In: Eurographics Conference, Vienna (2006)

43. Hildenbrand, D., Lange, H., Stock, F., Koch, A.: Efficient inverse kinematics algorithm based on conformal geometric algebra using reconfigurable hardware. In: GRAPP Conference, Madeira (2008)
44. Hildenbrand, D., Pitt, J.: The Gaalop home page. http://www.gaalop.de (2008)
45. Hocking, J., Young, G.: Topology. Addison-Wesley/Dover, New York (1961)
46. Hunt, J.: Vorticity and vortex dynamics in complex turbulent flows. Proc. CANCAM Trans. Can. Soc. Mech. Eng. **11**, 21 (1987)
47. Jeong, J., Hussain, F.: On the identification of a vortex. J. Fluid Mech. **285**, 69–94 (1995)
48. Kenwright, D., Henze, C., Levit, C.: Feature extraction of separation and attachment lines. IEEE Trans. Vis. Comput. Graph. **5**(2), 135–144 (1999)
49. Kimmel, R., Malladi, R., Sochen, N.: Images as embedded maps and minimal surfaces: movies, color, texture, and volumetric medical images. Int. J. Comput. Vis. **39**(2), 111–129 (2000). doi:10.1023/A:1008171026419
50. Koenderink, J.J.: The structure of images. Biol. Cybern. **50**, 363–370 (1984)
51. Kornprobst, P., Deriche, R., Aubert, G.: Non-linear operators in image restoration. In: Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition (CVPR '97), San Juan, 17–19 June 1997, p. 325. IEEE Computer Society, Washington, DC (1997)
52. Lindeberg, T.: Scale-Space Theory in Computer Vision. Kluwer Academic, Dordrecht (1994)
53. Löffelmann, H., Doleisch, H., Gröller, E.: Visualizing dynamical systems near critical points. In: Spring Conference on Computer Graphics and Its Applications, Budmerice, pp. 175–184 (1998)
54. Mann, S., Rockwood, A.: Computing singularities of 3D vector fields with geometric algebra. In: Proceedings of the IEEE Visualization, Boston, pp. 283–289 (2002)
55. Mattiussi, C.: The geometry of time-stepping. In: Teixeira, F.L. (ed.) Geometric Methods in Computational Electromagnetics. PIER, vol. 32, pp. 123–149. EMW, Cambridge (2001)
56. McCormick, B., DeFanti, T., Brown, M.: Visualization in scientific computing-a synopsis. IEEE Comput. Graph. Appl. **7**(7), 61–70 (1987). doi:10.1109/MCG.1987.277014
57. Naeve, A., Rockwood, A.: Course 53 geometric algebra. In: SIGGRAPH Conference, Los Angeles (2001)
58. Nielsen, M., Florack, L., Deriche, R.: Regularization, scale-space and edge detection filters. J. Math. Imaging Vis. **7**(4), 291–308 (1997)
59. Perona, P., Malik, J.: Scale-space and edge detection using anisotropic diffusion. IEEE Trans. Pattern Anal. Mach. Intell. **12**(7), 629–639 (1990)
60. Perwass, C.: The CLU home page. http://www.clucalc.info (2005)
61. Perwass, C.: Geometric Algebra with Applications in Engineering. Springer, Berlin (2009)
62. Petsche, H.-J.: The Grassmann bicentennial conference home page. /http://www.uni-potsdam.de/,u/philosophie/grassmann/Papers.htm (2009)
63. Pham, M.T., Tachibana, K., Hitzer, E.M.S., Yoshikawa, T., Furuhashi, T.: Classification and clustering of spatial patterns with geometric algebra. In: AGACSE Conference, Leipzig (2008)
64. Preußer, T., Rumpf, M.: Anisotropic nonlinear diffusion in flow visualization. In: Proceedings of the Conference on Visualization '99: Celebrating Ten Years. IEEE Visualization, San Francisco, pp. 325–332. IEEE Computer Society, Los Alamitos (1999)
65. Reyes-Lozano, L., Medioni, G., Bayro-Corrochano, E.: Registration of 3d points using geometric algebra and tensor voting. J. Comput. Vis. **75**(3), 351–369 (2007)
66. Rosenhahn, B., Sommer, G.: Pose estimation in conformal geometric algebra. J. Math. Imaging Vis. **22**, 27–70 (2005)
67. Rudin, L., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. Physica D **60**, 259–268 (1992)
68. Sapiro, G.: Geometric Partial Differential Equations and Image Analysis. Cambridge University Press, Cambridge (2001)
69. Sapiro, G., Ringach, D.L.: Anisotropic diffusion of multi-valued images with applications to color filtering. IEEE Trans. Image Process. **5**(11), 1582–1585 (1996)
70. Stalling, D., Steinke, T.: Visualization of vector fields in quantum chemistry. Technical report, ZIB m-96–01 (1996)

71. The homepage of geomerics ltd. http://www.geomerics.com. Last visited 2015
72. Theisel, H., Weinkauf, T., Hege, H.-C., Seidel, H.-P.: Saddle connectors – an approach to visualizing the topological skeleton of complex 3D vector fields. In: Proceedings of the IEEE Visualization, Seattle, pp. 225–232 (2003)
73. Theisel, H., Weinkauf, T., Hege, H.-C., Seidel, H.-P.: Grid-independent detection of closed stream lines in 2D vector fields. In: Proceedings of the Vision, Modeling and Visualization 2004, Standford, 16–18 Nov 2004, pp. 421–428. http://www.courant.nyu.edu/~weinkauf/publications/bibtex/theise104b.bib
74. Tomasi, C., Manduchi, R.: Bilateral filtering for gray and color images. In: Proceedings of the Sixth International Conference on Computer Vision (ICCV), Bombay, 04–07 Jan 1998, p. 839. IEEE Computer Society, Washington, DC (1998)
75. Tonti, E.: The reason for analogies between physical theories. Appl. Math. Model. **1**(1), 37–50 (1976/1977)
76. Treinish, L.A.: Data explorer data model. http://www.research.ibm.com/people/l/lloydt/dm/dx/dx_dm.htm (1997)
77. Tschumperlé, D.: Fast anisotropic smoothing of multi-valued images using curvature- reserving PDE's. Int. J. Comput. Vis. **68**(1), 65–82 (2006). ISSN:0920–5691
78. Tschumperlé, D., Deriche, R.: Vector-valued image regularization with PDE's: a common framework for different applications. IEEE Trans. Pattern Anal. Mach. Intell. **27**(4), 506–517 (2005)
79. Veldhuizen, T.: Using C++ template metaprograms. C++ Rep. **7**(4), 36–43 (1995). Reprinted in C++ Gems, ed. Stanley Lippman
80. Vemuri, B.C., Chen, Y., Rao, M., McGraw, T., Wang, Z., Mareci, T.: Fiber tract mapping from diffusion tensor MRI. In: Proceedings of the IEEE Workshop on Variational and Level Set Methods (VLSM'01), Vancouver, 13 July 2001, p. 81. IEEE Computer Society, Washington, DC (2001)
81. Venkataraman, S., Benger, W., Long, A., Byungil Jeong, L.R.: Visualizing Hurricane Katrina – large data management, rendering and display challenges. In: GRAPHITE 2006, Kuala Lumpur, 29 Nov–2 Dec 2006
82. Weickert, J.: Anisotropic Diffusion in Image Processing. Teubner-Verlag, Stuttgart (1998)
83. Weickert, J.: Coherence-enhancing diffusion of colour images. Image Vis. Comput. **17**, 199–210 (1999)
84. Weinkauf, T.: Extraction of topological structures in 2D and 3D vector fields. PhD thesis, University Magdeburg. http://tinoweinkauf.net/ (2008)
85. Weinkauf, T., Theisel, H., Hege, H.-C., Seidel, H.-P.: Boundary switch connectors for topological visualization of complex 3D vector fields. In: Data Visualization 2004. Proceedings of the VisSym 2004, Konstanz, 19–21 May 2004, pp. 183–192. http://www.courant.nyu.edu/~weinkauf/publications/bibtex/weinkauf04a.bib
86. Zomorodian, A.J.: Topology for Computing. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, Cambridge/New York (2005). First published 2005, reprinted 2009

Index

Symbols
2D-Fourier transform, 18
3D Mesh Bilateral Filter Definitions, 1655–1656

A
Abell 1689 ISOCAM data, 2094
ABEMML algorithm, 422–423
ABMART algorithm, 421–422
Absorbing energy estimation, 582–583
Absorption coefficient reconstruction, 583–584
Absorption potential, 949–950
Acoustic stress-confiment condition, 1085–1086
Acoustically homogeneous media, 1121–1122
Active arrays, 1298–1299
Adams, John Couch, 9
Additive noise model, 1354–1357
Adjoint, 781, 788
Affine low-rank minimization, 227–229
Algebraic reconstruction technique (ART), 374–377, 818–822, 1003
Ambiguity function, 784
Ambrosio and Tortorelli phase-field elliptic approximations, 1551
Amplitude contrast model, 945, 979
Analysis operator, 1720
Anisotropic diffusion, 1617
Anisotropic diffusion PDE's, 2149–2157
Anisotropic total variation models, 1482–1487
Anomaly detection
  electrical impedance tomography, 537–545
  numerical methods, 541–544
  ultrasound imaging, 545–555
Aperture problem, 1614, 1963–1964
Approximate inverse method, 999–1001
Approximation
  error approach, 1060, 1065–1067
  Mumford and Shah functional, 1550–1551
Arc-length parameterization, 1787
Aristotle, 6
Assignment approach, 1950–1956
Asymptotic behavior, neighborhood filters, 1619–1622
Atkinson–Wilcox expansion, 656
Augmented Lagrangian method, 1524
Autoconvolution equations, 116–118
Autocorrelation, 772
Autocovariance function (ACF), 1378, 1379
Automated image registration (AIR), 152

B
Backprojection, 782
Bacon, Francis, 8
Bag of features (BOF), 1901
Balls phantom, 1016–1020
Banach spaces, 98–104, 274
Banach's theorem, 1719
Band extraction, 2080
Band-limited signals, 19–20
Bandwidth, 776
Basis, 1721–1722
Basis functions, 813
Bayes estimation, 329–330
Bayes' formula, 1369
Bayes risk, 1441
Bayes' rule, 314
Bayesian approach, 2086
Bayesian estimate, 818
Bayesian framework
  approximation error approach, 1065–1067
  for inverse problem, 1061–1062
  inference, 1062
  likelihood and prior models, 1062–1063
  nonstationary problems, 1063–1065
Bayesian inference, level set based tracking, 1932–1933
Beer–Lambert law, 1035, 1036

© Springer Science+Business Media New York 2015

O. Scherzer (ed.), *Handbook of Mathematical Methods in Imaging*,
DOI 10.1007/978-1-4939-0790-8

Bellman optimality principle, 1874
Beltrami equation, 728
Bernoulli matrices, 221–222
Besov norms, 1010
Bessel bound, 1720
Bessel function, 1146
Bessel sequence, 1720–1721
Best k-term approximation, 213–214
Bilateral filter, 1602, 1603, 1647–1658
Blagovestchenskii identity, 1226, 1227
Blind deconvolution, 52, 1360–1361, 2089–2090
Blob, 814
Block iterative ART, 1004
Bojarski identity, 686
Boltzmann–Shannon entropy, 270, 282
Born approximation, 661, 685–686, 769, 793–794, 1048–1049, 1179–1181
  error estimate, 1258–1261
  explicit formula for slab, 1255–1258
  modified, 1194
Bouguer's law, 14
Boundary control method, 1207–1208, 1211–1234, 1273–1275
Boundary derivative operator, 1045
Boundary distance function, 1209
Boundary rigidity problem, 1209
Bregman iteration, 1521–1524
Burg's entropy, 365

C
Calderón's technique, 717
Canonical dual frame, 1723
Cardinal B-splines, 1677, 1681
Cartesian products, 1346
Cauchy data, 720–721
Cauchy–Green deformation tensor, 1831
Cauchy strain tensor, 2013
Cauchy stress law, 1827
Caustics and trapped rays, 1270–1271
C-band, 765
Census transform, 1983
Centroidal Voronoi sampling, 1871
Centroidal Voronoi tessellation (CVT), 1871
Chains, 2122–2124
Chambolle's algorithm, 1473–1474
Chambolle's dual method, 1516–1518
Chan-Golub-Mulet's primal-dual method, 1516
Chan–Vese model, 483
Charged Coupled Device (CCD), 963
χ^s-Divergences, 138–140
Chunking, 1446
Cimmino–Landweber methods, 374–377
Circular integrating detectors, 1125
Closed-form inversion formulas, 1149–1151
Coarea formula, 1461
Coarse-to-fine strategy, 1969
Cochains, 2124–2128
Coherence, 220–221, 782
Coherent interferometric (CINT) imaging
  active arrays, 1331–1335
  cross-correlations, 1321
  mean point spread function, 1324–1326
  SNR, 1322–1324, 1326–1327
  Wigner transform, 1327–1331
Coherent multi-valued geometry, 2151
Cohomology, 2128–2130
Collimated source model, 1042–1043
Color image recognition, 1427–1428
Color image restoration, 1577–1579
Color level set technique
  binary level set model, 484
  of multiple shapes, 484
  for reservoir characterization, 486–487
  for tumor detection, 485–489
Combined phase and amplitude contrast model, 945
Complete electrode model (CEM), 711, 746
Complete orthonormal system, 23
Complex geometrical optics (CGO), 717–720, 724
Complimentary slackness conditions, 288
Compressive sensing (CS), 206–248, 1992–1994
Computational anatomy, 1762
Computational efficiency, Geometric Algebra using Gaalop, 2139–2141
Computerized tomography (CT), 232, 802, 804, 805
Conditional covariance, 1062
Conditional density, 1347
Conditionally Gaussian hypermodels, 1361–1362
Conditional mean (CM), 1062, 1366–1368
Conditional probability, 1061
Conductivity distribution, 489
Cone beam CT, 867–868
Conformal Geometric Algebra, 2136–2139
Congruences, 1864
Conical tilt electron microscope tomography, microlocal analysis, 889–891
Conical tilt ET, 868–871
Conjugate gradients, 837
Continuity results, X-ray transform, 863–864

Continuous case
acceptable preferred data, 400–401
missing data, 401–402
preferred data selection, 401
Continuous case with Y = h(X)
censored exponential data, 403–405
example, 402–403
general approach, 405
Continuous-to-discrete (C-D) imaging models, 1105–1106
Contrast transfer function (CTF), 961
Convergence properties, 238–240
Convex analysis, 281–293
Convex Mean Value Theorem, 278
Convolution operator, 346
Convolution theorem, 18, 1727
Correlation, 772
Correlation coefficient, 131
Correspondence problem, 1947–1949
Cost function, 490, 1440
Co-tangential space, 2104
Covariance operator, 1840
Co-vectors, 2104–2105
Covering number, 1443
Cowpea mosaic virus (CPMV), 1020
Crack detection, 479
Cramér–Rao lower bound, 1364
Cross-entropy (CE) problems, 363–366
Cross-range resolution, 785–786
Cryo-fixation method, 942
Csiszár Divergences, 137–141
Curvature-preserving PDE's, 2152–2156
Curvelet decomposition, 1210–1214, 1239
Cut locus distance function, 1216

D
Data, NGC 2997, 2065
Data access ordering, 818, 832
Data acquisition geometry, 971–973
Data-collection manifold, 776
Data compatibility, 101
Data discrepancy, 976–977
Data model, 2102
Dr. David E. Kuhl, 907
DC-programming, 1980
Deblending, 2059, 2078
Deblurring, 1349–1350
Decoder, 224
Decomposition methods, 675–677
Deconvolution, 50–52, 266–268, 2081–2094
Delaunay mesh, 1872
Denoising, 1349, 1607–1614, 1656
Density estimation, 133–137
Derivative-of-Gaussian (DoG) filters, 1954
Detector model, 982
Deterministic pixel measure, 127
Dichotomy class divergences, 139
Dielectric medium, 1172
Diffeomorphism, 736, 1777–1780, 1798–1800
Difference imaging, 747
Differentiable manifold, 1864
Differential forms, 2119–2122
Differential motion approach, 1950–1956
Differential motion estimation, 1951–1954
Differential path length factor (DPF), 1036
Diffuse source model, 1043
Diffusion approximation (DA)
Boltzmann hierarchy approach, 1040
boundary conditions, 1041–1042
collimated source model, 1042
Monte Carlo diffusion approach, 1044
numerical solution methods, 1043
photon density and photon current, 1040
validity, 1043
Diffusion-based optical tomography (DOT), 1047
Diffusion geometry, 1866–1868
Diffusion kernel, 1866–1868
Diffusion tensors images, 1513–1514
Digital difference analyzer (DDA), 821
Dijkstra's algorithm, 1874–1875
Dirac delta distribution, 502
Direct regularization methods, 93–107
Direct scattering problems, 652–653
Helmholtz equation, 654–657
Maxwell equations, 661–666
obstacle scattering, 657–660
Dirichlet boundary value problems, 1222
Dirichlet-to-Neumann map, 114, 703, 720–728, 732–734, 739–741
Discrepancy principle, 435
Discrete Fourier transform (DFT), 1735
Discrete geodesics, 1846–1848
Discrete histogram, 131
Discrete modulation, 1735
Discrete path length, 1845
Discrete Picard condition, 57
Discrete reconstruction problem, 816
Discrete-to-discrete (D-D) imaging models, 1108–1109
Discrete total variation, 1477–1479, 1505–1507
Discretized Laplace–Beltrami operator, 1880
Distance functions, 1949
Distorted wave Born approximation, 957
Domain inpainting, 1508
Dose problem, 977–978

Down-range resolution, 785
Dual and primal-dual methods, 1515–1516
Dual decomposition, 1891
Duality gap, 289
Duality operator, 1774
Dual problem, 1439
Dyadic decomposition, 1240
Dynamic imaging, 1053
Dynamical shape priors, variational segmentation, 1935–1936
Dyson equation, 1049

E
Edge linking model, 1485
Effective dipole method, 543
Efficient ordering, 832
Eikonal equation, 1235–1236, 1875–1876
Einstein ring, 11
Ekeland's variational principle, 291
Elasticity-based PCA, 1837–1843
Elastic registration, 149
Elastic regularization functionals, 2012–2014
Elastic scattering, 947
Elastic shape average, 1835–1837
Electrical impedance tomography
  bibliography and open questions, 544–545
  mathematical model, 538–539
  numerical methods, anomaly detection, 541–544
  physical principles, 537–538
  voltage perturbations, asymptotic analysis, 539–541
Electromagnetic waves scattering, 766
Electron Λ-tomography (ELT), 994–998
Electron micrograph, 805
Electron microscope tomography (ET), 853–856
Electron microscopy, 805
  EM algorithm, 339–341
  macromolecular assemblies, 337
  maximum likelihood problem, 337–339
  weighted least-squares problem, 342
Electron–specimen interaction, 946–947
Electron wave, 947
Electrosensing, 702
Electrostatic potential models, 950
Elliptic equation, 114–116
Embedded manifolds, 1865
Emission computed tomography (ECT), 185
Emission tomography, 330–337
  Gibbs smoothing, 347–348
  Good's roughness penalization, 345–347
  Poisson random variable, 331–332
  regularization, need for, 342–343
  Shepp–Vardi EM algorithm, 332–337
  smoothed EM algorithms, 343–344
Empirical risk minimization (ERM), 1442
Encoder, 224
Energy function, 1963
Entropy regularization, 1006–1008
Epipolar constraint, 1974
Euclidean geometry, 1864
Euler equations, 345
Euler-Lagrange equations, 1552, 1559, 1566
Euler–Poincaré equation, 1776–1777
Evolution equations, 1802
Exact penalization veto, 101
Expectation–maximization (EM) algorithms, 460–464, 1371–1374
  ART and Cimmino–Landweber methods, 374–377
  continuous case, 399–405
  convergence, 426
  Csiszar and Tusnady approach, 409–412
  data binning, 323–327
  deconvolution problem, 319–327
  discrete case, 397–398
  empirical Bayes estimation, 329–330
  finite mixture problems, 423–426
  finite mixtures, 407
  Gibbs smoothing, 347–348
  Good's roughness penalization, 345–347
  and Kullback-Leibler distance, 407–409
  MART and SMART methods, 377–380
  maximum-likelihood problem, 311, 317
  minimum cross-entropy problems, 363–366
  missing data, 398
  monotonicity properties, *see* Monotonicity properties
  multinomial distribution, 414
  multinomial example, 406
  multiplicative iterative algorithms, 369–370
  multiplicative smoothing, 344
  non-negative solutions, for linear equations, 417–423
  nonnegative least squares, 366–369
  NSEM, 391–393
  ordered subset EM algorithm, 371–374
  Poisson sums, 412
  Poisson sums ECT reconstruction problem, 415–416
  Poisson sums using KL distance, 417
  Radon-Nikodym derivatives, 312
  row-action and block-iterative EM algorithms, 380–384

Shepp–Vardi EM algorithm, 332–337
smoothing operator, 343
stochastic EM algorithm, 393–396
Explicit vs. implicit shape representation, 1911–1913
Exponential charts, 1768
Extended Kalman filtering (EKF), 1064
Extension, movies, 1614
Exterior algebra, 2108
Exterior product, 2107
Extrapolation of band-limited signal, 19–20

F
Factorization method, 594–595, 654, 692–693
conducting obstacles, 603–610
crack problem, 614–615
half space problem, 612–614
impedance tomography, insulating inclusions, 595–603
inverse acoustic scattering, sound-soft obstacle, 616–622
inverse electromagnetic scattering, inhomogeneous medium, 622–627
local data, 611–612
Fan beam parameterization, 868
Far field approximation, 1179–1181
Far field operator, 269, 682, 686–688
Far field pattern, 653
Farthest point sampling (FPS), 1870–1871
Fast marching methods, 1876
f-Divergences, 137–141
Feature-based methods, 1899–1902
Feature extraction, 1394
Fenchel conjugate, 260, 281–284, 1439
Féjèr monotonicity, 349
Fenchel duality, 284–285
Fermat's rule, 277
Fermi–Dirac entropy, 282
Fiber-bundle classification scheme, 2114
Fick's law, 1040–1042
Field properties, 2116–2117
Figure of merit (FOM), 827
Filter, 787
Filtered backprojection (FBP), 782, 787, 811, 864–865, 906, 990–994, 1149
Filtering, 2078
f-Information, 141–146
Finite element method (FEM), 744, 1043
Finite mixture problems
acceptable data, 424
convergence of Mix-EM algorithm, 426
likelihood function, 423
Mix-EM algorithm, 425–426
motivating illustration, 424
probability density functions, 423
Finite time response operator, 1213
First order Born approximation, 956
First order pseudo-differential operator, 709
Fisher information matrix, 1358–1359
Fixed point theory, 292–293
Fleishman's bilateral filter, 1650
FLIRT, 151
Fluence rate perturbations, 584
Focus of expansion (FOE), 1959–1960
Footprint, 779, 780
FORBILD thorax phantom, 839
Formal adjoint, 781
Forward mapping, 1134–1135
Forward model, 944–971
Forward operator, 474, 491, 966–970
Fourier integral operators, 880–882
Fourier transform, 346, 1092, 1093, 1724–1726
Fréchet derivative, 500
Fréchet distance, 1789
Fragment, 2119
Frame, 1722–1724
Frame bounds, 1722–1724
Frame coefficients, 1723
Frame decomposition, 1723
Frame operator, 1723
Fredholm integral equations, 269–271, 297–298
Free surface boundary condition, 1039
Frequency
bands, 765
domain, nonlinear problem in, 1265–1266
shifts, 1727
Functional photoacoustic tomography, 1088–1089
Function spaces, 859–860
Fundamental solution, 766
Fuzzy C-means, 1650

G
Gabor analysis, 1718
Bessel sequences, Hilbert spaces, 1720–1721
convolution, involution and reflection, 1727
discrete gabor systems, 1734–1738
Fourier transform, 1724–1726
frame bounds, 1722
frame decomposition, 1723
non-separable atoms, 1751–1752
orthonormal basis, 1721–1722
pseudo-inverse operator, 1719–1720
sampled STFT, 1746–1747

Gabor analysis (*cont.*)
STFT, 1727–1730
tight frames, 1724
time-frequency shifts, 1731, 1733
translation and modulation, 1726–1727
Gabor atom, 1732
Gabor frame theory
finite discrete periodic signals, 1735–1736
in $\ell^2(\mathbb{Z})$, 1734–1735
in $\mathbb{C}^L$, 1736–1738
in $L^2(\mathbb{R}^d)$, 1730–1734
Galilei, Galileo, 7–8
Gaussian beams, 1230–1234
Gaussian low-pass filter, 1954
Gaussian matrices, 221
Gaussian noise, 1356
Gaussian point spread function, 17, 19
Gauss map, 128
Gauss-Newton method, 68, 69
Gelfand widths, 210–211, 224–227
Generalized cross validation (GCV), 66
Generalized Kaiser-Bessel window function, 814
Generalized Kullback–Leibler (KL) divergence, 2085
Generalized multidimensional (GMDS) scaling, 1889–1891
Generalized portrait algorithm, 1398
Geodesic active contour model, 1484–1486
Geodesic equation, 1766–1767
Geodesic path, 1829
Geodesics, 1865
Geometric Algebra, 2109–2110, 2132
Geometrical optics, 954–955, 1237–1238
Geometric realization, 1872
Global minimizers, 170
Global point signature (GPS), 1888
Global stability, 730–731
Golub-Kahan bidiagonalization (GKB), 62
Gradient descent evolution, kernel density estimator, 1929–1930
Gradient direction, 501
Gradient methods, 435
Grady segmentation method, 1635
Graph cut methods, 1524
Green's function, 767, 1045–1046, 1048, 1049, 1152
Green's operator, 1055
Green-St.-Venant strain tensor, 2013
Grenander's theory of deformable templates, 1762
Gromov–Hausdorff distance, 1822, 1888–1891
Gromov–Wasserstein distance, 1892
Group testing, 211

H
Hahn–Banach/Fenchel duality circle, 275
Hahn–Banach theorem, 275–276
Half-time reconstruction problem, 1102–1104
Hamilton flow, 1244
Hamiltonian approach, 1772–1773
Hamilton–Jacobi-type equation, 497
Hammersley–Clifford theorem, 1353
Hand-written digit recognition, 1425–1427
Hankel transform, 1146
Hard margin classifier
linear learning, 1400–1402
nonlinear learning, 1413–1414
Hard margin regression nonlinear learning, 1414–1415
Hausdorff distance, 1789, 1883–1884
Hausdorff measure, 1548
Hausdorff topology, 2112
Heating function, 1084
Heat kernel, 1867
Heat kernel signature (HKS), 1900
Heat operator, 1868
Heaviside function, 499
Helical CT, 805
Helmholtz equation, 654–657, 1147
Henyey–Greenstein scattering function, 1038, 1040
Herglotz wave function, 677
Hessian matrix, 81
High-range-resolution (HRR) pulse, 773
Hilbert spaces, 96–98, 104–106, 1763
Hinge loss function, 1403
Histogram, 2119
concentration, 1635–1639
Hoffman phantom, 924
Hölder regularity, 1467–1468
Holder stability, 1133–1135
Homeomorphism, 1217–1221
Homogeneous coordinates, 1956
Homology, 2128–2130
Horizontality condition, 1772
Horn and Schunck approach, 1964–1967
Huygens' principle, 659, 677–682
Hybrid method, 678
Hyperelastic regularization functionals, 2014–2015
Hyperelastic regularizer, 2015

I
Illumination, 945–946
Image, 50
deblurring, 50–52
denoising, 266–268

dose, 983
formation, 1349
kernel, 784
mode, TEM, 941–942
motion analysis, 1946
processing, 1676, 1701, 1704, 1705
reconstruction, 472–585, 811, 904
registration, 126, 148–149
restoration, 1488–1495
restoration with impulsive noise, 1573–1577
segmentation, 1511–1513, 1910
segmentation via Bayesian inference, 1913–1915
Image quality (IQ) phantom, 923
Impediography
bibliography and open questions, 565
mathematical model, 563–564
physical principles, 561–562
substitution algorithm, 564–565
Implicit shapes
dynamical shape priors, 1931–1936
linear dynamical models, 1934
Impulse response, 784
Incident field, 767
Inelastic electron scattering, 949–952
Inexact matching, 1810–1811
Inexact Newton methods, 73
Infimal convolution, 283
Infrared thermal imaging
asymptotic analysis, temperature perturbations, 555–557
bibliography and open questions, 560
numerical methods, 549–554
physical principles, 555
Inhomogeneous coordinates, 1956
Inhomogeneous medium, scattering, 660–661, 669–671
Initial-boundary value (IBV) problem, 1126
Initial convex sequence estimator (ICSE), 1381
Initial monotone sequence estimator (IMSE), 1381
Initial positive sequence estimator (IPSE), 1380
Inpainting, 1349
Integrated autocorrelation time (IACT), 1380
Integrated data function, 1092
Intensity-based image registration technique, 150
Intensity operator, 962
Interpolation by radial basis function, 1417–1418
Interscale relation, 2077
Intrinsic alignment
invariance, 1923–1924
translation invariance, 1924–1925
Intrinsic symmetry, 1897–1898
Invariance, 1947–1949
Invariant shape similarity, 1861, 1882–1883
canonical forms, 1885–1888
graph-based methods, 1891
Gromov–Hausdorff distance, 1888–1891
Gromov–Wasserstein distances, 1892–1893
rigid similarity, 1883–1885
shape DNA, 1893
Inverse, 811
Inverse attenuated Radon transform (IART), 906, 910–911, 917–918
Inverse crime, 1383
Inverse gamma distribution, 1361
Inverse kinematic problem, 1209
Inverse medium problem, Newton iterations, 682–683
Inverse problems, 977–985
Inverse Radon transform, 906, 913–917
Inverse scattering, 268–269, 296–297, 1275
iterative and decomposition methods, 672–686
problem, 652
qualitative methods, 686–696
uniqueness, 667–671
Inverse synthetic-aperture radar (ISAR), 774–775
resolution, 784–785
Involution, 1727
Isometric embedding, 1869
Isometry, 1863
Isotropic undecimated wavelet transform (IUWT), 2062
Iteration operator, 313, 353
Iterative algebraic techniques, 1153
Iterative closest point (ICP) algorithms, 1884
Iterative image reconstruction, 1110–1113
Iteratively regularized Gauss–Newton method, 451–455
Iteratively reweighted least squares (IRLS), 236–243
Iterative methods, 1001–1005, 1471–1477
Iterative regularization, 37–39, 61, 435
Iterative step, 818
ITK, 151

J
Jacobian matrix, 747
Jensen-Shannon divergence, 138
Joint restoration and segmentation, 1582–1584

K
Ka-band, 765
Kaczmarz method
in frequency domain, 1266
time domain, 1261–1263
Kaczmarz method time domain, 1261–1263
Kaczmarz's algorithm, 32
Kaczmarz-type methods, 458–460
Kalman filtering, 1064
Karcher mean, 1803
Kendall's theory, 1760
Kernel density estimation, 134–135
level set domain, 1926–1929
Kernel-trick, 1412–1413
Kirchhoff imaging, 554
Kirchhoff/physical optics approximation, 659
Kirchhoff–Poisson formulas, 1143
Kirchhoff-type imaging, broad range of frequencies, 551–552
Kriging, 1415–1419
Ku-band, 765
Kuhn-Tucker conditions, 1438
Kullback–Leibler divergence, 138, 309–310, 315

L
Lagrange multipliers, 289–291, 1437
Lagrangian duality, 287–289
Landweber iteration, 39
Landweber method, 61
Laplace–Beltrami operator, 734, 1865
Laplacian field, 110
Laplacian kernels, 1771
Laplacian operator, 1083
Large deformation diffeomorphic metric mapping (LDDM), 1013
Large deformation diffeomorphic metric matching (LDDMM) approach, 1987
Laser-based photoacoustic tomography, 1086–1087
Layered medium, 1193–1194
L-band, 765
L-curve, 67
Least squares cost functionals, 500–501
Least squares methods, inverse medium problem, 683–685
Lebesgue-space, 1821
Lens-less imaging, 960
Lens rigidity problem, 1210
Level set methods, 1851
binary media, 479–481
color level set, *see* Color level set technique
cost functionals, 490
cracks/thin shapes, 487–489
Eulerian derivatives, 492–493
geometric quantities, 495–497
material derivative method, 493–494
piecewise constant level set function, 482–483
single smooth level set function, 482
transformations and velocity flows, 491–492
vector level set, 483
Levenberg–Marquardt method, 447–451
Likelihood density, 1348
Likelihood distribution, 1061
Limited angle Lambda CT, 867
Limited data X-ray CT, 887–888
Linear approximations, 1052–1053
Linear detectors, 1124–1125
Linear diffusion, 1634–1639
Linear discriminants, 1398
Linear Gaussian shape priors, 1916–1920
Linear hard margin support vector classifier, 1398
Linear inverse problems, 262–263, 293–295
alternating projection theorem, 31–32
band-limited signals, 19–20
compact operators and SVD, 27–30
Cormack's inverse problem, 14–16
deblurring problem, 17–19
discretization, 39–43
forward and reverse diffusion, 16–17
iterative regularization, 37–39
linear operators, 25–27
Moore–Penrose inverse, 30–31
PET, 20–21
Platonic inverse problem, 11–14
renaissance, 7
Tikhonov regularization, 32–37
weak convergence, 23–25
Linear learning, 1398–1410
Linear least squares approach, 1396
Linear sampling method, 593, 628–631, 653, 688–691
Lippmann–Schwinger equation, 660, 1178
frequency domain, 768
integral equation, 654, 768
Lipschitz properties, 272–273
Lipschitz stability, 1133–1135
Lloyd–Max algorithm, 1871
Local energy decay estimates, 1132
Local minimizers, 168–170
Logistic loss functions, 1407

M
Magnetic induction tomography (MIT), 706
Magnetic resonance elastography
  asymptotic analysis, displacement fields, 575–577
  bibliography and open questions, 579–580
  mathematical model, 573–575
  numerical methods, 578–579
  physical principles, 573
Magneto-acoustic imaging
  magnetic resonance elastography, 573–580
  magneto-acousto-electrical tomography, 567–570
  photo-acoustic imaging of small absorbers, 580–584
  with magnetic induction, 570–572
Magneto-acousto-electrical tomography
  mathematical model, 567–568
  physical principles, 567–570
  substitution algorithm, 568–570
Majorization-minimization (MM), 1530–1532
Marginal densities, 1346
Marginal polytope, 1986
Markov chain, 1376
Markov Chain Monte Carlo (MCMC) sampling, 1374–1382
Markov model, 1351
Markov random field (MRF), 1353
Matched filter, 770–772
Material derivative method, 493–494
Mathematical shape theory, 1760
MATLAB code, 75, 77
Matrix-valued total variation, 1505
Matsushita's Divergences, 140–142
Max formula, 278
Maximal flow methods, 1477–1482
Maximal violating pair strategy, 1448
Maximum a posteriori (MAP) estimate, 751, 2086
Maximum a posteriori (MAP) estimation, 1062, 1366
Maximum entropy method, 1006
Maximum likelihood (ML) method, 2083
Maximum likelihood and Fisher information, 1358–1359
Maximum likelihood estimation, 306–309
Maxwell's equations, 661–666, 1172
  in Fourier domain, 1173–1174
  initial conditions of, 1174–1175
Mean curvature motion (MCM), 1656–1657
Measurement matrix, 214
Measurement vector, 810
Membership functions, 1896
Mercer kernels, 1412
Mercer's theorem, 1430–1431
Merit function, 490
Metric discretization
  diffusion distance, 1880–1882
  Dijkstra's algorithm, 1874–1875
  eikonal equation, 1875–1876
  implicit surfaces and point clouds, 1880
  metrication errors and sampling theorem, 1875
  parallel marching, 1878–1880
  parametric surfaces, 1878
  triangular meshes, 1876–1878
Metric distortion, 1768
Metric spaces, 1863
  diffusion distances, 1868
  diffusion geometry, 1866–1868
  Euclidean geometry, 1864
  isometries, 1863
  Riemannian geometry, 1864–1866
  topological spaces, 1863
Metrication error, 1875
Microlocal analysis, 871–883
Microlocal analysis, X-ray CT, 884–887
Microlocal regularity principle, 995
Microwave breast screening, 475–477, 523–525
Middlebury database, 1962
Migration imaging expectation, 1316–1317
Migration imaging SNR, 1318–1319
Modeling error, 1383
Modica-Mortola theorem, 1551
Modulation operator, 1734
Modulation transfer function, 1493–1494
Moment conditions, 1142
Moment generating function (MGF), 1058
Momentum map, 1775
Momentum representation, 1767
Monge-Kantorovich formulation, 1989
Monotone operator, 292
Monotonicity method, 743
Monotonicity properties
  Féjèr monotonicity, 349
  function inequality, 353
  Gibbs smoothing, 359–362
  Shepp–Vardi EM algorithm, 350–352
  smoothed EM algorithm, 354–359
Monte Carlo diffusion approach, 1044
Morozov principle, 1007
Morozov's discrepancy principle, 34
Morphological measures, 128–129
Motion compensated filter, 1614
Motion field, 1951
MR image reconstruction, 199

Multi-frame blind deconvolution (MFBD), 52, 78–81
Multi-scale finite element approximation, 1852–1853
Multi-slice method, 956
Multi-task-learning, 1423
Multichannel total variation, 1504
Multidimensional scaling (MDS), 1887
Multiple anomalies detection, 543–544, 559–560
Multiplicative art algorithm (MART), 377–380
Multiplicative noise removal, 195–196
Multipolar fluids, 1828
Multiresolution analysis, 1676, 1677, 1679, 1681, 1685–1690
Multiresolution support, 2077
Multiscale vision model, 2075, 2076
Multivariate Gaussian random variable, 817
Mumford-Shah functionals, 498
MUSIC, 631–635
MUSIC-type algorithm, 543–544
MUSIC-type imaging, single frequency, 549–550

N

Nachman's method, 754
Nash embedding theorem, 1866
Near-infrared (NIR) frequency, 1089
Neighborhood filter/sigma filter, 1600, 1647
NEMA NU 4-2008 phantom, 928
Neumann function, 539
Newton iterations, inverse obstacle scattering, 672–675
Newtonian fluid, 1827
Newton's method, 1805–1806
Newton type methods, 447–458
NL-means algorithm, 1609–1616
Noise, 771
Nonasymptotic bounds, 170–172
Non-blind restoration, 1570
Non-convex functionals, 1979–1981
Non-convexity of $\mathcal{S}^{\text{quad}}$, 2014
Non-convex model, 1420
Nonconvex regularization, 172–178
Nonlinear elasticity model, 1841
Non-linear image registration
  distance functionals, 2010–2012
  ill-posedness and regularization, 2012
  image, 2008–2009
  mathematical setting, 2007–2008
  transformation, 2008–2010
  variational formulation, 2008
Nonlinear Landweber regularization, 435–442
Nonlinear learning, 1410–1428
Nonlinear maximization, 1370–1371
Nonlinear problem frequency domain, 1265
Nonlinear smoothing operator, 344
Nonlinear statistical shape priors, 1918–1920
Nonlocal Mumford-Shah regularizers, 1584–1592
Nonlocal total variation, 1507
Non-negative solutions, for linear equations
  acceleration, 418–422
  general case, 418
  regularization, 418
  using prior bounds on, 420–423
Non-physical scattering transform, 756
Non-rigid volumetric objects, 1827–1834
Non-smooth convex functionals, 1977–1979
Nonsmooth regularization, 178–185
Nonstationary inverse problems, 1063–1065
Nonstationary Neumann-to-Dirichlet map, 1207
Non-stochastic EM algorithm (NSEM)
  continuous case, 391–393
  discrete case, 393
Non-topological representations, 2118–2119
Non-trapping condition, 1131
Non-uniqueness set, 1127–1130
Nonlinear problem frequency domain, 1265
Normal cone mapping, 267
Normal velocity, 496
Notions of resolution, 1015
Nuclear magnetic resonance (NMR) imaging, 207
Nuclear norm minimization, 227
Nuisance parameters, 981
Null hypothesis, 829
Null space property, 216–217
Nyquist-Shannon sampling theorem, 1952

O

Object classification, 2059
Object reconstruction, 2092
Obstacle scattering, 657–660, 667–671
Ogden materials, 2015
One-step late (OSL), 1374
Online estimation methods, 1995
Ontological scheme and seven-level hierarchy, 2113–2116
Optical coherence tomography (OCT), 1171
  forward operator of, 1181–1185
  frequency domain, 1171–1172
  full field, 1171
  measurements of, 1175–1177
  polarization-sensitive, 1172

standard, 1171
time domain, 1171–1172
Optical imaging
adjoint field method, 1054–1055
diffusion approximation, *see* Diffusion approximation (DA)
error models, construction of, 1071–1072
experiment and measurement parameters, 1067–1070
FEM meshes and discretization accuracy, 1070–1072
forward mapping, 1046–1047
light propagation and probabilistic interpretation, 1056–1059
linearization, 1052–1053
MAP estimates, 1072–1074
perturbation analysis, 1048–1049
prior model, 1069–1071
radiative transfer equation, 1037–1039
Robin to Neumann map, 1046
Schrödinger form, 1047–1048
spectroscopic measurements, 1035–1036
Optical tomography
Bayesian framework, *see* Bayesian framework
image reconstruction, 1059
Optics, 957–962
Optimal control formulation, 1811–1812
Optimal current patterns, 713
Optimal transport, 1988–1991
Ordered subset EM algorithm (OSEM), 371–374
Order of approximation, 1686
Orthogonality condition, 1142
Orthonormal set, 23
Overcomplete frame, 1724

P
Parabolic scaling, 1240
Parallel beam geometry, 972
Parallel beam transform, 869
Parallel marching, 1878
Parameter distribution, 482
Parametric maximum flow algorithm, 1481
Parametric shape representations, 1915–1920
Parametrix approach, 1153–1155
Parametrix-type reconstructions, 1158
Parseval's identity, 23
Parseval's theorem, 1867–1868
Partial similarity, 1893–1896
Partial symmetry, 1898
Path-based shape spaces, 1822
Path-based viscous dissipation, 1827–1831
Path optimization, 1807
PDE models and local smoothing filters, 1614–1618
Perona-Malik model, 1618, 1625
Persistent radar, 795
Perturbation analysis
Born approximation, 1048–1049
Rytov approximation, 1049–1051
Petroleum engineering, 477–479, 525–527
Phase field approach, 1851–1852
PhaseLift algorithm, 230
Phase retrieval problem, 227, 229–230, 987–988
Photo-acoustic imaging, small absorbers
bibliography and open questions, 585
mathematical model, 580–581
physical principles, 580
reconstruction algorithms, 581–584
Photoacoustic tomography (PAT), 1118–1163
acoustic heterogeneities, 1103–1104
data redundancies, 1102
discrete imaging models, 1104–1106
finite-dimensional object representations, 1107–1108
finite transducer bandwidth, 1096
Fourier-Shell identity, 1093–1094
frequency-dependent acoustic attenuation, 1099–1100
functional PAT, 1088–1089
laser-based PAT, 1086–1088
non-point-like transducers, 1097–1099
RF-based PAT, 1087–1088
speed-of-sound distribution, 1100–1102
thermoacoustic effect and signal generation, 1083–1086
thermoacoustic tomography, *see* Photoacoustic tomography (PAT)
universal backprojection algorithm, 1092–1093
Photometry, 2059
Photon measurement density function (PMDF), 1053
Picard's criterion, 29
Picture distance measure, 826
Picture function, 810
Piecewise-constant mumford and Shah segmentation, 1558–1561
Piecewise-smooth Mumford and Shah segmentation, 1561–1567
Pixel, 813
Pixelization, 965
Planar detectors, 1123–1125
Point detectors, 1120
Point source method, 676

Point-spread function (PSF), 51, 784, 2094
Pointwise perturbations, 548
Poisson–Kirchhoff formula, 1121
Poisson likelihood, 1374
Poisson noise, 2070
Poisson process, 1372
Polar Cone Calculus, 286
Polar format algorithm (PFA), 787
Polarization tensor, 543
Polarization tensor properties, 541
Polyharmonic B-splines, 1694–1695, 1701–1705
Polynomial kernel, 1433
Pontryagin maximum principle (PMP) theorem, 1773
Positron emission tomography (PET), 20–21, 51, 909, 2011
  maximum likelihood problem, 349
  Shepp–Vardi EM algorithm, 332–337
Potential function (PF), 159, 172–173
Potential method, 675
Precision, 1362
Pressure perturbations, 546
Primal-dual active-set method, 1519–1521
Primal dual algorithm, 233–236
Primal-dual approaches, 1475–1477
Primal-dual hybrid gradient method, 1518–1519
Principal component analysis, 1821
Prior, 817
Prior density, 1348
Probabilistic models, 1994–1995
Probability density, 1375
Probability density function (PDF), 309, 817, 1058
Probability distribution, 1346–1348
Probe method, 637–639
Projection approximation, 956
Projection matrix, 816
Projection method, 822
Proximal mapping, 284
Pseudodifferential operators, 875–880
Pseudo-inverse operator, 1719
Pshenichnyi–Rockafellar Conditions, 287
Puri-Vincze divergences, 140
Pythagorean theorem, 7, 30

Q
Quadratic programming, 1527–1528
Quantum mechanical models, 951
Quantum mechanics, 970
Quasiphotons, 1231
Quotient spaces, 1783–1784

R
Radar imaging, 232
Rademacher's theorem, 164
Radial basis functions (RBF), 1654
Radiative transfer equation (RTE), 1037–1039
RAdio Detection And Ranging (Radar), 764
  SAR, *see* Synthetic aperture radar (SAR) imaging
Radon line transform, 860–863
Radon, Johann, 9
Radon transform, 777, 810
Radon's inversion formula, 803
Random media, imaging
  in cluttered media, 1303
  in forward model, 1283–1288
  in least squares inversion, 1292
  in long range scaling and Gaussian statistics, 1308–1309
  in normal operator, 1294–1295
  in passive arrays, 1295–1298
  in random model, 1303–1305
  in robustness to additive noise, 1299–1301
  in setup for imaging, 1313–1314
  in statistical moments, 1309–1313
  in time reversal process, 1292–1294
  in wave propagation, 1306–1307
Random partial Fourier matrices, 222–223
Range
  alignment, 788–791
  resolution, 785
Range alignment, ISAR, 788–791
Ray optics, 958–959
Ray transform, 944
Read-out noise, 965
Reciprocity, 710
Reciprocity gap function, 691
Reconstruction algorithm, 810
Reconstruction formula for OCT, 1185–1202
  for a dispersive layered medium with focused illumination, 1193–1196
  for a dispersive medium, 1189–1193
  for a non-dispersive medium, 1188
  for a non-dispersive medium with focused illumination, 1188–1189
  for an anisotropic medium, 1196–1202
Reconstruction methods, 988–1014
Reconstruction problem, 742–757
Reconstruction problem, electron tomography, 974–977
Reconstruction procedure, 756

Redundant frame, 1724
Reflectors, 1271–1273
Region competition segmentation, 1634
Region of interest (ROI) data, 888
Regression, 1395
Regression correction, neighborhood filter, 1625–1628
Regularization Concave on $\mathbb{R}_+$, 197–199
Regularization parameters, 66–67
Regularization scheme, 751
Regularization theory, 32–34
Regularized iterative non-linear methods, 749–753
Reinforcement learning, 1395
Relative entropy, 142
Relativistic corrections, 949
Relaxation parameter, 819, 836
Rellich's lemma, 656, 663
Reproducing kernel Hilbert spaces (RKHSs), 1415–1419, 1428–1436, 1764–1765
Resolution
  cross-range, 785–786
  range, 785
Restoration of noisy signal, 183
Restricted isometry property (RIP), 217–220
Richardson-Lucy (RL) algorithm, 2087–2089
Ridge regression, 1397, 1408
Riemannian geometry
  embedded manifolds, 1865
  geodesics, 1865
  n-dimensional manifold, 1864
  Riemannian metric tensor, 1865
  rigidity, 1866
Riemannian manifolds, 1206, 1211–1214, 1774, 1826–1827, 1865
Riemannian metric, 1209, 1765
Riemannian metric tensor, 1865
Riemannian shape space, 1823
Riemannian submersion, 1769–1770
Riesz basis, 1678
Riesz–Fredholm theory, 658
Riesz potential, 787
Riesz representation theorem, 24, 1763
Right-inverse operator, 1719
Rigid similarity, 1883–1885
Rigid symmetry, 1897
Robin boundary condition, 1043
Robin to Dirichlet map, 1046
Robust linear programming (RLP), 1420
ROI tomography, 866–867
Rollnick condition, 952
Root-mean-square error (RMSE), 1657
Rotating targets, 775
Row-action method, 821
Rytov approximation, 1049–1051

S
Sampling and probe methods, 742
Sampling theory, 1011
Sandwich theorem, 278
SAR imaging, 891–895
Scalar curvature equation, 1221
Scale-invariant heat kernel signatures (SI-HKS), 1901
Scaled gradient projection (SGP), 2088
Scaling function, 1677, 1679, 1683
Scattered field model, 770
Scattering operator, 948
Scattering phase function, 1038
Scattering relation, 1238–1240
Schrödinger equation, 718–720
Schrödinger form, 1047–1048
Second dyadic decomposition, 1240–1242
Second generation starlet transform algorithm, 2066–2068
Second-order cone programming (SOCP), 1529–1530
Segmentation problem, 1540
Selection rules, 1007
Self-similarity and symmetry
  intrinsic symmetry, 1897
  partial symmetry, 1898
  repeating regular structure, 1899
  rigid symmetry, 1897
  spectral symmetry, 1897–1898
Semi-blind restoration, 1570–1572
Semi-smooth newton's method, 1519
Sensitivity functions, 1053–1054
Separatrices, 2147–2148
Sequential discrepancy principle, 100
Series expansion method, 813
Shannon entropy, 142
Shape derivative, 495
Shape discretization
  implicit surfaces, 1873
  parametric surfaces, 1873
  sampling, 1869–1870
  simplicial complexes, 1872–1873
Shape distances, level sets, 1922–1923
Shape distribution, 1884
Shape DNA, 1893
Shape evolution
  calculus of variations, 500–505
  Eulerian derivatives, 492–493
  geometric constraints, 497–499

Shape evolution (*cont.*)
gradient direction, 501
Heaviside function, 507–512
least squares cost functionals, 500–501
material derivative method, 493–494
rough level set functions, 514
and shape optimization, 516–518
simple shapes and parameterized velocities, 516
smooth velocity fields, 515
smoothed level set updates, 512–514
of thin shapes, 504–505
TM-waves, 503–504
transformations and velocity flows, 491–492
velocity field, 502–503
Shape identification problems, 592
Shape sensitivity analysis, 492–493
min-max principle, 506–507
TM-waves, 506
Short-time Fourier transform (STFT), 1727–1730
efficient Gabor expansion, 1746–1747
modulation priority, 1747–1751
translation priority, 1748, 1751
visualization, 1747–1751
Shot noise, 964–965
Shrinkage estimators, 192
Shrinking, 1448
Shunt model, 704
Signal-to-noise ratio (SNR), 771, 1071
Silver–Muller finiteness conditions, 663
Silver–Muller radiation conditions, 662
Similarity filters, 1653–1655
Simultaneous algebraic reconstruction technique (SART), 1004
Simultaneous iterative reconstruction technique, 1004
Simultaneous multiplicative algebraic reconstruction technique (SMART), 377–380
Single anomaly detection, 542–543, 557–558
Single frequency
backpropagation-type imaging, 550–551
MUSIC-type imaging, 549–550
Single photon emission computed tomography (SPECT), 909, 1372
Singular sources method, 635–637
Singular support and wavefront set, 871–874
Singular value decomposition (SVD), 28, 57–59
Skeleton level, 2117
Skin depth, 707
Slater condition, 291
Small angle approximation, 955
Small-angle case
cross-range resolution, 785–786
down-range resolution, 785
Small-scene approximation, 772–773
Smooth approximations, 1245
Smoothness and support conditions, 1142
SNARK09, 822
Sobolev functions, 1542
Sobolev metrics, 1794
Sobolev spaces, 1212
Soft field imaging, 710
Soft margin classifier nonlinear learning, 1402–1404, 1414
Soft margin regression linear learning, 1404–1407
Soft margin regression nonlinear learning, 1415
Sommerfeld's finiteness condition, 655
Sonography, *see* Ultrasound imaging
Source conditions, 434, 456
Space-variant restoration, 1579–1582
Sparse approximation, 210
Sparse modeling, 2069
Sparsity, 213–214
Sparsity data model, 2061
Sparsity promoting regularization, 1011–1013
Spatio-temporal approach, 1973
Spearman's rank correlation coefficient, 151
Spectral analysis, 912
Spectral factorization, 59
Spectral symmetry, 1897–1898
Spectral theorem, 1429–1430
Spherical Bessel function, 1142
Spherical mean operator, 1121–1122
Spiral CT, 805, 839
Spline interpolation problem, 1764
Spline reconstruction technique (SRT)
PET, 918–922
SPECT, 930–933
Split Bregman iteration, 1523–1524
Splitting methods, 1532–1534
Spotlight SAR, 780
Standard phase contrast model, 967–968, 979
Star-galaxy separation, 2059
Staring radar, 795
Starlet reconstruction, 2064–2066
Starlet transform, 2056
Starlet wavelet transform, 2062
State-based approach, 1826
State-based elastic deformation, 1829–1833
vs. path-based dissimilarity measure, 1833–1834
State estimation framework, 1064

State space representation, 1064
Static relative permittivity, 475–477
Statistical distance measures, 130–146
Statistical hypothesis testing, 826
Statistical methods
  additive noise model, 1354–1357
  counting noise, 1357
  hierarchical models, 1359–1362
  informative/noninformative priors, 1359
  maximum likelihood and Fisher information, 1358–1359
  maximum likelihood and maximum a posteriori estimation, 1363–1366
Statistical significance, 827
Steepest descent and minimal error method, 446
Stochastic EM algorithms, 313
  acceptable data, 396
  conventional formulation, 394
  E-step and M-step, 393–394
  incorrect proof, 395–396
Stochastic filtering, 1995
Stopping rules, 434–435
Stratton–Chu formula, 662
Streak line, 2142
Structure determination problem, 939
Structure tensor, 1969
Subdifferential, 275–276
  subdifferential sum rule, 279
Sublinear function, 274
Sum-of-squared-difference (SSD), 2011
Superiorization methodology, 822
Superresolution, 1509–1511
Surface impedance, lower bounds, 693–695
SUSAN filter, 1602
Synthesis operator, 1720
Synthetic aperture radar (SAR)
  spotlight, 780
Synthetic aperture radar (SAR) imaging, 779–782, 856–858
  applications, radar imaging, 792–796
  historical background, 764–766
  mathematical analysis of methods survey, 774–786
  mathematical modeling, 766–773
  numerical methods, 787–791
  radar frequency bands, 765
  spotlight SAR, 780–782
  stripmap SAR, 781–782
  unmodeled motion, problems related to, 792–793
  unmodeled scattering physics, problems related to, 792–796
System matrix, 1108

T
Tangent space, 1864
Tangent vector, 1864
Tangential divergence, 494–495
Tangential vectors, 2104
Tartaglia, Niccolò, 7
Temporal coherence of silhouettes, 1932
Tensors, 2105–2107
TE-waves, 507
Thermoacoustic tomography (TAT), 1118–1163
Tight frame, 1722
Tikhonov regularization, 32–38, 96–98
Tikhonov regularized solution, 1363
Time-correlation single photon counting (TCSPC) systems, 1036
Time-frequency lattice, 1731
Time-frequency shifts (TFshifts), 1726
Time line, 2142
Time-reversal imaging, 552–554
Time shift, 1726
Tobacco mosaic virus (TMV), 1020
Tomographic transforms properties, 859–871
Tomography, 801–842
Tomosynthesis, 53–55, 81–85
Topological derivatives, 520–522
Topological realization, 1872
Topological skeletons, 2117–2118
Topological space, 1863
Topology, 2130–2132
Total variation (TV), 1503–1507
  denoising problem, 266
  functionals, 498
  regularization, 161, 172, 1008–1011
Trace-based PDE's, 2152
Traditional data model, 2057
Transfer admittance, 710
Transform method, 811
Transform pairs construction, 911–912
Transition function, 1864
Transitive group action, 1784–1786
Translation and scale invariance via alignment, 1925–1926
Translation operator, 1726
Transmission eigenvalues, 695–696
Transmission electron microscope (TEM), 807, 940–943
Transport equation, 1237–1238
Transport theory, 1037
Transportation map, 1989
Trapping condition, 1131
Triangle inequality, 1863
Triangular meshes, 1872, 1876–1878
Trilateral filters, 1652–1653

Turntable geometry, 775
TV denoising problem, 1463–1468
Two-dimensional case, 1623–1625

U
Ultrasound imaging, anomaly detection
  bibliography and open questions, 554–555
  frequency domain, asymptotic formulas, 545–546
  numerical methods, 549–554
  physical principles, 545
  time domain, asymptotic formulas, 547–548
Ultrasound imaging, by wave equation, 1253–1275
Unbounded linear operators, 26
Unconditional basis, 1721
Unification, mathematical systems, 2134
Uniform handling, different geometric primitives, 2135–2136
Uniform uncertainty principle, 211
Unique continuation principle, 660
Uniqueness of reconstruction, 1135–1136
Uniqueness set, 1127
Universal backprojection algorithm, 1092–1093
Universal backprojection formula, 1149

V
Vanderbilt database, 151
Variable projection method, 70
Variational image registration, 1987
Variational image restoration, 1567–1569
Variational inequalities, 106–107
Variational methods, 1005–1013
Variational principles, 291–292
Variational regularization, 59–60, 98–104
Vector and fiber bundles, 2110–2111
Vector field visualization, 2141–2149
Vector level set, 483
Vector-valued case, 1628–1632
Vertical bundle, 1780
Viscosity solution, 1876
Viscous fluid shape space, 1843–1849
Viscous moment tensor, 577
Visibility condition, 1139–1140
Visible (audible) singularities, 1136–1137
Visualization, 805
Voltage perturbations, asymptotic analysis, 539–541
Volterra iteration scheme, 1247
Voronoi regions, 1870

W
Walking person, 1930–1931
Wasserstein distances, 1885
Wave equation, 766–767
Wave equation model, 1119–1120
Wavelet denoising, 2073
Wave packets, 1210–1214
Weak-scattering/single scattering approximation, 769
Weighted back-projection (WBP), 813, 990–994
Weighted boundary measurements perturbations, 547
Weil–Peterson metric, 1797
Well-posed problem, 4
Wentzel–Kramers–Brillouin (WKB) approximation, 954–955
Weyl–Heisenberg frame, 1731
White Gaussian noise Removal, 178, 192–194

X
X-band, 765
X-ray tomography (CT), 849–853

Y
Yosida approximation, 284

Z
Zero-boundary condition, 1041
Zero level set, 480